



---

---

**Statistical Policy  
Working Paper 17**

**Survey Coverage**

---

---

**Prepared by  
Subcommittee on Survey Coverage  
Federal Committee on Statistical Methodology**

**Statistical Policy Office  
Office of Information and Regulatory Affairs  
Office of Management and Budget**

**May 1990**

**MEMBERS OF THE FEDERAL COMMITTEE ON  
STATISTICAL METHODOLOGY**

**(April 1990)**

**Maria E. Gonzalez (Chair)  
Office of Management and Budget**

**Yvonne M. Bishop  
Energy Information  
Administration**

**Warren L. Buckler  
Social Security Administration**

**Charles E. Caudill  
National Agricultural  
Statistical Service**

**John E. Cremeans  
Office of Business Analysis**

**Zahava D. Doering  
Smithsonian Institution**

**Joseph K. Garrett  
Bureau of the Census**

**Robert M. Groves  
Bureau of the Census**

**C. Terry Ireland  
National Computer Security  
Center**

**Charles D. Jones  
Bureau of the Census**

**Daniel Kasprzyk  
Bureau of the Census**

**Daniel Melnick  
National Science Foundation**

**Robert P. Parker  
Bureau of Economic Analysis**

**David A. Pierce  
Federal Reserve Board**

**Thomas J. Plewes  
Bureau of Labor Statistics**

**Wesley L. Schaible  
Bureau of Labor Statistics**

**Fritz J. Scheuren  
Internal Revenue Service**

**Monroe G. Sirken  
National Center for Health  
Statistics**

**Robert D. Tortora  
Bureau of the Census**

## PREFACE

The Federal Committee on Statistical Methodology was organized by the Office of Management and Budget (OMB) in 1975 to investigate methodological issues in Federal statistics. Members of the committee, selected by OMB on the basis of their individual expertise and interest in statistical methods, serve in their personal capacity rather than as agency representatives. The committee conducts its work through subcommittees that are organized to study particular issues and that are open to any Federal employee who wishes to participate in the studies. Working papers are prepared by the subcommittee members and reflect only their individual and collective ideas.

The Subcommittee on Survey Coverage studied the survey errors that can seriously bias sample survey data because of undercoverage of certain subpopulations or because of overcoverage of other subpopulations. The purpose of this report is to heighten the awareness of survey planners and data users regarding the existence and effects of coverage error, and to provide survey researchers with information to evaluate the trade-offs between coverage error and survey costs. The report profiles selected methods for controlling and measuring the effects of coverage errors using examples from Federal sampling frames and surveys. The report includes seven case studies based on Federal surveys that illustrate selected aspects of coverage errors.

The Subcommittee on Survey Coverage was cochaired by Cathryn S. Dippo of the Bureau of Labor Statistics, Department of Labor, and Gary M. Shapiro of the Bureau of the Census, Department of Commerce.

**MEMBERS OF THE SUBCOMMITTEE ON  
SURVEY COVERAGE**

Cathryn S. Dippo (Co-chair)  
Bureau of Labor Statistics (Labor)

Gary M. Shapiro (Co-chair)  
Bureau of the Census (Commerce)

Raymond R. Bosecker  
National Agricultural Statistics Service (Agriculture)

Vicki Huggins  
Bureau of the Census (Commerce)

Roy Kass  
Energy Information Administration (Energy)

Gary L. Kusch  
Bureau of the Census (Commerce)

Melanie Martindale  
Defense Manpower Data Center (Defense)

D.E.B. Potter  
Agency for Health Care Policy and Research (Health and Human Services)



## ACKNOWLEDGMENTS

This report is the result of the collective work and many meetings of the Subcommittee on Survey Coverage. All of the subcommittee members made significant contributions to the text of the report, taking responsibility for various sections of the report during the long period of preparation.

All of the members of the Federal Committee on Statistical Methodology reviewed several drafts and made many important suggestions. The subcommittee wishes to recognize in particular the valuable contributions made by the following committee members: Yvonne Bishop, Joseph Garrett, Charles Jones, Daniel Kasprzyk, Fritz Scheuren, Monroe Sirken, and Robert Tortora. The subcommittee also benefited significantly from an outside review of the final draft by Steven Heeringa and Benjamin Tepping.

The subcommittee also thanks the following persons: John Paletta and Richard Pratt for preparing the Current Population Survey and Producer Price Index case studies, respectively; Robert Casady and Charles Cowan for contributing to the section on sample design strategies; and Rosalie Epstein of the Bureau of Labor Statistics for editing the report.

# TABLE OF CONTENTS

	Page
LIST OF TABLES . . . . .	vii
LIST OF FIGURES . . . . .	viii
EXECUTIVE SUMMARY . . . . .	1
CHAPTER 1. Coverage errors occurring before sample selection . . . . .	3
1.1 Conceptual or relevance error . . . . .	4
1.2 Frame construction and maintenance . . . . .	8
1.2.1. Classification of frame errors . . . . .	13
Missing elements; clusters of elements appearing on list; blanks or foreign elements; duplicate elements; incorrect auxiliary information	
1.2.2. Frame maintenance . . . . .	15
New frame elements; inactive frame elements; misclassified elements; out-of-scope elements; split-out or combined frame elements	
1.2.3. Match-merging of independent source lists . . . . .	21
1.3. Sample design strategies to minimize coverage error . . . . .	22
1.3.1. Defining target population to equal frame population . . . . .	23
1.3.2. Random-digit dialing sampling . . . . .	23
1.3.3. Multiple frame sampling . . . . .	24
1.3.4. Sampling rare populations . . . . .	25
1.3.5. Estimation procedures . . . . .	27
1.4. Evaluation methods . . . . .	28
1.4.1. Macro-level analysis . . . . .	28
1.4.2. Micro-level analysis . . . . .	29
CHAPTER 2. Coverage errors occurring after initial sample selection . . . . .	31
2.1. Incorrect association of frame with reporting unit(s) . . . . .	31
2.1.1. Location errors . . . . .	31
2.1.2. Classification errors . . . . .	33

2.1.3. Temporal errors . . . . .	36
2.2. Listing errors . . . . .	38
2.2.1. Area segment listing errors . . . . .	39
Studies measuring error; an alternative to area listing	
2.2.2. Household listing errors . . . . .	43
Motivational causes; lack of correspondence between survey designer's and respondent's residency concepts; effect of household listing errors; methods for reducing household listing errors	
2.2.3. Nonhousehold listing errors . . . . .	47
2.3. Other nonsampling errors . . . . .	47
2.3.1. Recording errors . . . . .	47
2.3.2. Responses from nonsampled units . . . . .	49
2.3.3. Coverage errors resulting from nonresponse . . . . .	50
CONCLUSION . . . . .	53
APPENDIX A. CASE STUDIES	
Introduction . . . . .	55
A.1. Annual Survey of Manufactures (ASM) . . . . .	56
A.2. National Long-term Care Survey (NLTCs) . . . . .	61
A.3. National Master Facility Inventory (NMFI) . . . . .	65
A.4. Producer Price Index (PPI) . . . . .	71
A.5. Quarterly Agricultural Surveys (QAS) . . . . .	77
A.6. Monthly Report of Industrial Natural Gas Deliveries . . . . .	83
A.7. Current Population Survey (CPS) . . . . .	89
APPENDIX B. GLOSSARY OF ACRONYMS . . . . .	96
APPENDIX C. GLOSSARY OF TERMS . . . . .	97
REFERENCES . . . . .	106

## LIST OF TABLES

Number	Title	Page
1.	Selected sampling frames used for Federal surveys . . . . .	10
2.	Scope of frame versus population of interest for selected surveys . . . . .	33
3.	Reinterview classification of units originally classified as noninterview: October 1966 . .	34
4.	Reinterview classification of units originally classified as noninterview: April to September 1966 . . . . .	35
5.	Reinterview classification of units originally classified as noninterview: 1987 . . . . .	35
6.	Type B rates for the Survey of Income and Program Participation and the Current Population Survey, 1985-87 (percent) . . . . .	35
7.	Selected surveys in which the frame sampling unit and the final sampling unit are the same . . . . .	38
8.	Selected surveys in which the frame sampling unit and the final sampling unit differ . . .	38
9.	Examples of surveys requiring field listing . . . . .	39
10.	Comparison of A.C. Nielsen 1982 field canvass of housing units with 1980 census housing unit counts by block group or enumeration district (National Nielsen Television Index Survey segments only) . . . . .	40
11.	Number of listing errors found in Labor Force Survey study (Statistics Canada) . . . . .	41
12.	Reasons units were added and deleted during reinterview, as determined by reconciliation--area segments only: October 1966 . . . . .	41
13.	Estimates of percent net CPS within-household undercoverage relative to the 1980 census for males aged 25 and over by their household status (standard errors in parentheses) . . . . .	45
14.	1986 average coverage ratios by age, sex, and race for CPS . . . . .	92
15.	1986 average coverage ratios for Hispanics by age and sex for CPS . . . . .	92

## LIST OF FIGURES

Number	Title	Page
1.	Typical physical flow of natural gas from gas well to industrial customer (custody relationship) . . . . .	84
2.	Possible financial flows (ownership) from gas well to industrial customer (equity relationship) . . . . .	84
3.	Industrial gas estimates from Form EIA-857 submissions: Total United States . . . . .	87

## EXECUTIVE SUMMARY

Coverage errors can cause serious biases in estimates based upon sample survey data. Undercoverage may be substantial in many surveys, especially of selected subpopulations. For example, the estimated undercoverage of Hispanic males aged 14 and over is 23 percent in the Current Population Survey (see appendix A.7). In economic surveys, new businesses may be missed at a higher rate than older ones. If the characteristics of the missed portion of the population are very different from those of the covered portion, serious biases in the survey estimates for the total population will result.

The purpose of this report is to heighten the awareness of survey program planners and data users concerning the existence and effects of coverage error and to provide survey researchers with information and guidance on how to assess and improve coverage in sample surveys. The report outlines the possible sources and effects of coverage error by documenting current knowledge of coverage errors in Federal surveys. It also profiles selected methods for controlling, measuring, determining the effects of, and reducing coverage errors using examples from Federal surveys and sampling frames.

This report utilizes a broad definition of coverage error. Some authors have included only errors associated with the sampling frame. Here, however, coverage error is defined to include all possible sources of error which are not classified as observational or content errors (U.S. Department of Commerce 1978b). For example, errors or mistakes leading to noncoverage of target population units (undercoverage), errors or mistakes leading to the inclusion of units which are not members of the target population (overcoverage), and failure to elicit a response for a sampled population unit (nonresponse) are included.

The report narrative is structured to follow the sequential procedures typically used in a survey. Other approaches, including one based upon a typology of sampling units (housing units, persons, and establishments), were considered but discarded because of the complexity of many surveys. (An excellent discussion of the coverage errors in housing unit surveys can be found in United Nations (1982).) The survey process has been divided, for the purpose of this report, into two components. Chapter 1 discusses coverage errors which might occur before the first stage of sampling. Issues associated with the creation and maintenance of sampling frames and the choice of sampling frame and strategy are included. Chapter 2 discusses coverage errors which might occur after the first-stage sampling units are selected. Coverage errors associated with field listing, screening, subsequent sampling operations, interviewing, and processing are presented, along with overcoverage due to volunteer respondents. Nonresponse as an important source of coverage error and bias, particularly in housing unit surveys and mail surveys of establishments, is also discussed.

Each chapter includes a detailed discussion of the circumstances leading to coverage errors. A discussion is also provided regarding the seriousness of the errors, their effects on survey estimates, and methods for controlling, measuring, and improving survey coverage. Numerous studies, which have been conducted to measure either overall frame coverage or the effects on coverage of selected data collection procedures, are cited throughout the report. One large single source of coverage error identified in this report is within-household listing of persons. In general, coverage error is a more significant problem in housing unit surveys than in establishment surveys.

Throughout the report, examples are used to illustrate, not to encompass, the diversity of knowledge and experience derived from surveys conducted by Federal agencies. Although the examples in the text are necessarily brief, a more detailed examination of selected coverage issues is provided in appendix A, which presents exemplary material from the following surveys: Annual Survey of Manufactures, National Long-term Care Survey, National Master Facility.

Inventory, Producer Price Index, Quarterly Agricultural Surveys, Monthly Report of Industrial Natural Gas Deliveries, and Current Population Survey. Readers are encouraged to compare their current knowledge and practices concerning coverage with those of other Federal agencies as represented by the examples in the report. To assist the reader, glossaries of acronyms and terms are included at the end of this report as appendix B and C.

## **CHAPTER 1**

### **COVERAGE ERRORS OCCURRING BEFORE SAMPLE SELECTION**

This chapter's goal is to provide a comprehensive set of evaluative tools which will enable users to identify and minimize potential coverage problems associated with a survey research program or specific research project, to assess the strengths and weaknesses of alternative research methodologies as these relate to potential survey coverage error, and to identify overly ambitious research projects and recast them into an achievable framework.

Four major subjects are discussed: Conceptual or relevance error, frame construction and maintenance, sample design strategies to minimize coverage error, and coverage evaluation methods. The chapter delineates the thinking, planning, and assessing processes which should precede and inform a complete survey design with its associated sampling plan.

The first section of the chapter contains a discussion on the importance of thinking carefully about prospective research and the necessity of using clear and concise language in the statement defining the research project or program. Attention to correct and clear thinking about, and specification of, research goals, concepts, and targeted population(s), an often neglected or abbreviated phase of research planning, helps to avoid or minimize many coverage problems at the outset.

Types of frame errors, standards for selecting or building a high-quality frame, and the many complex issues associated with correct and thorough frame maintenance, including match-merging independent source lists for updating and correcting frames, are discussed in the section on frame construction and maintenance. Not only are the major and minor problems arising from the failure to maintain frames illustrated with many examples, but evaluative criteria by which potential users of already existing frames may assess the appropriateness and adaptability of these frames for their own surveys are provided. The goal of this section is to provide the tools by which to identify appropriate existing frames, to assess those frames, and to determine when either supplemental or additional frames may be needed, or when a totally new frame must be built.

The third section presents sample design strategies that can minimize coverage errors associated with specific frame weaknesses. Moreover, design strategies for sampling rare populations for which existing frames are incomplete or inefficient are discussed. The section closes with a discussion of estimation procedures which compensate for known coverage error in the sampling frame(s).

Both macro- and micro-level analysis, as methods for measuring frame coverage, are discussed in the last section of the chapter. The degree of coverage error is measured routinely in many Federal establishment surveys. For example, reconciliations are made at the Bureau of the Census between economic census totals and corresponding totals in the Current Industrial Reports annual survey for census years to measure and improve coverage. Similarly, the National Agricultural Statistics Service (NASS) conducts a continuous survey program for the agriculture sector and compares inventory and production estimates with those obtained in the Census of Agriculture. Administrative data are also used to measure coverage of establishment surveys. For example, the Bureau of Labor Statistics makes annual comparisons between the employment reported to State unemployment insurance systems and establishment employment estimates from the monthly Current Employment Survey.

For the housing unit surveys conducted by the Bureau of the Census, a demographic approach is used to estimate the degree of coverage error. This approach is similar to what is termed demographic analysis, where the coverage of the decennial census rather than of survey data is



analyzed using other sources. However, using census data as a benchmark for survey coverage must be done cautiously, since the coverage error detected supplements the coverage error that already exists in the census results.

### 1.1. Conceptual or relevance error

Coverage errors can be caused by incorrect specifications of the concepts to be measured or the population(s) to be targeted by the survey. Incorrect specifications often result from conceptual errors. Some of these are hasty or incomplete thinking concerning the goals of the research, faulty reasoning or incorrect assumptions about the measurable characteristics of targeted groups, and failure to recognize existing information as relevant (or irrelevant) to the specifications being written. Incorrect specifications can sometimes be spotted by their vague, nonspecific, or ambiguous language. These faulty specifications, in turn, can lead to the construction of incomplete, inadequate, or otherwise flawed frames.

Hansen, Hurwitz, and Pritzker (1967) present the general concept of the mean-square error of a survey estimate which includes a term for the "relevance of the survey specifications as related to the requirements." This is the squared difference between a statistic which constitutes the ideal goals of the statistical survey and a statistic based upon the specifications actually set for the survey if carried out precisely according to defined goals. Using vague or ambiguous language in terms of the ideal goals can lead to greater relevance error because this language can increase any difference between ideal goals and actual specifications. Thus, relevance error is a type of coverage error, since failure to specify correctly the concepts to be measured can lead to the construction of a flawed frame.

To ensure a more useful and complete frame, a clear, precise statement of the research question(s) and population(s) of interest needs to be written down, with careful attention to the exact language used. This is particularly important when the ideal goals are proposed by nontechnical sponsors or appear in enabling legislation. It may even be useful to write down what is not being researched and who is not being targeted, especially if exclusions may be of interest to some client or to researchers generally. Taking the time to think through the possible meanings of key terms and variables and, if needed, to determine whether and how the population(s) and concepts have been defined and researched by others can reduce duplication of effort, reveal previous conceptual errors, and highlight potential frame construction problems. Even in a recurring survey, a review of the concepts and definitions can be very useful. For example, this effort may reveal changes in how the target population(s) and concepts are being defined by other researchers.

Sometimes, it may even be necessary to devise new terminology or revise definitions rather than to perpetuate the use of terms which now seem too general or otherwise objectionable. For instance, the "black" population of the United States has not always been called "black," and it may soon be preferable to use "African American." A review of shifts over time in race and ethnic concepts used in Federal research reveals various intersecting but nonidentical definitions for the black population, such as "nonwhite" and "colored." Such examples show how the use of vague terms makes it difficult to know who or what has been studied and how, over time, changes in terminology have generally been made to increase detail or specificity (and, thus, measurement accuracy), even at some cost in data continuity.

The language used to denote concepts, key variables, and the like should aim for concreteness and clarity. General equivalents of concepts (such as "education" to stand for "years of school completed") or of populations (such as "children" to stand for "persons aged 17 and under") should be avoided. Although very difficult, making the attempt to write specifications using "standard scientific terminology" rather than the "language of everyday life" wherever possible should help one to avoid vagueness in defining populations and concepts. Vague definitions of

populations and concepts tend to create coverage errors because they lead to inappropriate unit inclusions on, or exclusions from, a frame and even to naming a population which cannot be adequately represented on a frame.

A good rule to follow in examining the initial formulation of a problem is to ask a series of questions:

- To what population(s) of units does this problem refer?

Distinguish among populations about which information is sought, those which will be frame units, and those which may be reporting units, if different from the frame units. For example, suppose one wished to do research on "the scholastic achievement (as measured by grades) of children of recent immigrants." In this case, "children of recent immigrants," more suitably specified perhaps as "persons aged roughly 5 to 17 enrolled in Grades 1 through 12 of the U.S. public schools and living in a household in which at least one related head has been resident in the United States 5 or fewer years," would be the population about which information is sought. However, it seems likely that one might need to construct two or more frames in order to reach this population. One of the frames might have U.S. public schools as units, while another might consist of residential addresses to be screened. In this example, reporting units might well consist of two groups, school recordkeepers and parents or guardians.

- Is (are) this (these) population(s) observable or potentially measurable? How?

Continuing from the example above, one can see that the suggested specification of "children of recent immigrants" takes account of some of the presumably unobservable "children of recent immigrants," such as those who may be homeless and those who may not be currently enrolled in school. Among recent immigrants, those who entered the country illegally may not be observable, as well as those who died following entry, leaving school-age dependents. Sources for obtaining U.S. public school and residential addresses might be lists from various agencies. Thinking through all possible categories of the populations of interest should reveal those subsets which cannot be measured or reached; those whose measurements (observation) might be achieved; and those which seem reachable with some existing or proposed methodology. Thus, the "children" may be reached by means of a housing unit survey, school survey, and/or institutional survey (hospitals, orphanages).

- Are there one or more subsets of this (these) population(s) which cannot be measured/observed in some way? What are these? Would they ever be measurable?

Continuing the example of "children of recent immigrants," some possibly unobservable components of the populations have already been mentioned. The potentially measurable components might be those which cannot be reached now but which might be reached using a methodology that may be prohibitively expensive, such as scanning all death certificates or other sources of information to identify deceased recent immigrants. Thus, it may be useful to distinguish the inherently unobservable from the practically unobservable components of populations of interest.

- Does time enter into the answer to one or more of the questions above, in the sense that the measurable population(s) may change or may have changed?

Continuing the example of "children of recent immigrants," one may find that a change in a legal boundary or definition can turn "internal migrants" to "recent immigrants" or vice versa. This would happen, for example, if Puerto Rico became a U.S. State, thus solving the problem of how technically to classify migrants to the mainland, who would become

"internal migrants." Such a change might force a redefinition of the size and location of the populations of interest.

- Have previous efforts been made to build a frame of this (these) population(s)? What problems were encountered in frame construction? Was one of these faulty conceptualizations? Which of these problems has been solved?

This series of questions focuses on the need to locate previous research, to attempt to contact those who designed and conducted the research, or to obtain procedural histories about it and to evaluate carefully the definitions and language used by others. An assessment of previous research often reveals use of frames built for other purposes by still earlier researchers, especially when the frames are very expensive to assemble. Information needed for adequate frames may now be available (such as improved school lists) due either to improvements in information processing or to changes in laws regarding availability of administrative data.

Answering this list of questions has several important goals. The first is to decrease the slippage between the conceptual population(s) of interest and the actual units to be included on frames. The second goal is to facilitate the correct use of language, so that what can and cannot be researched is clearly understood. The third goal is to facilitate the specification of comprehensive and correct rules for frame maintenance. The fourth important goal is to help insure that the population(s) and concept(s) of interest will be defined and measurable, so that one can answer the research question(s) of interest with the greatest accuracy and completeness possible, with minimum coverage error.

In beginning to answer the questions given above (and there may be other useful ones to ask), background preparation might include a brief review of some of the literature on conceptualization. While extended treatments of this topic abound in the philosophy of science literature, statisticians such as Deming (1961) and applied researchers like Blalock (1968) have discussed the importance of conceptualization in more accessible and measurement-oriented works. Deming and others have discussed at length the "true value" of any variable or concept we attempt to measure, noting that there is no inherent "true value" and that the entity that we call the "true value" is a unique outcome of the concepts, assumptions, definitions, and procedures we use to arrive at it. With somewhat more focus on language, Blalock emphasizes the distinction between "concepts by intuition" and "concepts by postulation." These two kinds of concepts, one kind more or less abstract and the other more or less concrete, are linked for research and measurement purposes by assumptions (for example, the assumption that "education" is adequately reflected in "years of school completed"). It is upon these assumptions that researchers sometimes founder, not least because the language used is "everyday language" (for example, "children," "education," "worker").

The language of everyday life frequently is not suited to scientific inquiry. But when research is formulated solely or substantially in everyday language, the population(s) and concepts of interest will be named by this language as well. It is up to the researcher to guard against such usage and to clarify, specify, and define key terms, concepts, and populations, so that adequate frames can be constructed and coverage errors minimized. To build on an earlier example, a research organization might begin preparations for a study which will answer the question of "how children of recent immigrants are faring in school." Such a description of the research question may suffice for press coverage or to quickly summarize the general thrust of the effort for family and friends, but it reveals very little else. What is a measurable child? What does "recent" mean? Who is an immigrant? What is the process of "faring in school?" Is the intention to study a process, an outcome, or a set of outcomes? What is a "school?" Not only does the vague language used tell very little, it actually militates against thinking in clear ways. For example, once the word "school" is used, we probably unwittingly think of our own unique

individual "school" experiences, thus creating a tendency to omit by assumption other possibilities for defining "school."

This is not to say that one cannot use ordinary words; indeed, there is no method by which anyone can transcend all the limitations, contradictions, rules, and assumptions which are language itself. It means that there is a distinction to be made between the use of any word for purposes such as casual conversation and for research purposes. The difference usually lies in the modifiers and extended definitions containing the specificity and detail required for classification and measurement. In regard to frame construction specifically, concepts and population(s) of interest should be defined in such a way as to be observable and measurable in regard to the research question(s). (See appendix A.2 for an example of a carefully-defined target population.)

Sometimes it is possible to work with potential survey sponsors in order to gain assistance in formulating the original research question(s). Such discussions often reveal incomplete thinking and allow the researcher to eliminate undoable or excessively costly projects or to modify those involving major frame construction problems. Even after research agendas are set, as with enabling legislation, a meeting with sponsors will reveal intent and can save time and trouble later on. For example, recently proposed legislation called for the establishment of a Consumer Price Index (CPI) for the "elderly," where "elderly" was defined as "all persons 62 and over." In the initial proposal, the definition was "all persons 62 and over and retired." In both definitions, the targeted units were persons, whereas the usual units for interview are "consumer units." Not only do the different definitions imply potentially different sampling frames (and thus different cost levels) but also different procedures for constructing a CPI. Given these problems and others, had this legislation been enacted, it would have been necessary to determine the intent of Congress and establish a working definition for constructing an "elderly" index, so that the resulting research would have provided the information desired.

Finally, in thinking through any research agenda, attention should be paid to exclusions. Some frame (and research) exclusions are recognized and noted in many areas of research. One example might be something like: "This research focuses on immigrants who entered the United States within the last 5 years as identifiable in census data. It does not cover persons missed by the census. Some illegal entrants are included, but not identifiable, in the census data."

However, many exclusions are not noted, partly because researchers do not specify their work precisely, but often because the exclusions (or their existence in the real world) do not occur, or are unknown, to them. In addition, there is some lag between the emergence of new social phenomena and their explicit recognition.

In order to identify exclusions, it is often necessary to examine hidden assumptions and biases about the world. Reexamining topics from the perspectives of several disciplines and actually going out into the field might well be part of this process. As the result of these kinds of efforts and in various other complex ways, new concepts and populations for research do emerge. An example of this is the "hidden" economy.

Despite evidence that some forms of economic activity were not being included in the national accounts or were not the focus of serious research, "official" statistics and economic researchers in general failed until fairly recently to acknowledge such things as bartering, illegal activities of various kinds, and economic phenomena that were associated with other kinds of economic systems. Once the "hidden" economy or some subset or version of it was actually named, then vaguely described, it became easier for people to begin to think in new ways about the workings of the economic world. Once this kind of "breakthrough" occurred, it became easier to "see" exclusions. Today much more work has been done to identify, define, and attempt to research facets of an economic world which was largely ignored by the statistical establishment in the

past. However, for the "hidden" economy, as for any newly emergent topic, this process is by no means complete because inherent (and predictable) problems surrounding the interface between conceptualization and measurement have not yet been resolved.

One example of research intended to address the lack of terminological specificity in work on the "hidden" economy is McDonald's (1984) examination of the charge that Bureau of Labor Statistics employment, price, and productivity indexes are significantly affected by unreported economic activity. McDonald asserts: "Establishing the existence of a subterranean economy ... does not necessarily prove that government statistics are invalid. To determine whether a particular government statistic is affected also requires careful consideration of the way data are gathered ... and the relation between economic activities that may be covered by the survey and those that are not.... many of the critics of government statistics have simply not taken this necessary step" (p. 4). After discussing the most narrow through the most broad ways in which the underground economy had been defined in the literature to date, McDonald examined the extent to which evidence on the underground economy under any of these definitions implied mismeasurement of concepts measured by BLS data series and found that the critics had not proved their case. The importance of this work lies in its attempt to delimit explicitly several crucial interfaces between conceptualization and measurement pertaining to a researchable subject whose accumulated literature exhibited a notable lack of conceptual rigor. McDonald not only provided a solid point of departure for further conceptual and quantitative work on the "hidden" economy, but also pointed out what he had not covered in his assessment.

Of course, it hardly needs mentioning that it takes some creative thinking and observational acuity to "see" and to figure out how to name various forms of formerly unsuspected or illegal economic activity, let alone measure the monetary influence of these activities. Despite this, published economic research should not fail to mention any exclusions of already recognized "hidden" economic activity, where appropriate to the topic, and to state something about the potential effect of such exclusions on the findings at hand, regardless of whether this effect is minimal, large, or unknown. Since "hidden" economic activity has been the subject of a great deal of attention in recent years, even a statement that one or another form of it is irrelevant to the research being reported will reflect a prudent and thoughtful research approach and may prevent certain predictable criticisms.

More generally, a discussion of exclusions should be included in all published research as a matter of course. It should no longer be acceptable to omit mention of subpopulations which cannot be included on a frame. Excluding them from mention might well insure that no future attention will be accorded them and could give the false impression that existing frames are adequate or that new frames may not be needed. Put simply, mentioning exclusions points the way to future research and places the reported research in the correctly limited context. As a start, it is essential that statistical studies begin with a more extensive interaction between subject-matter experts and research methodologists. The gains can be large and may well enable researchers to avoid many of the other coverage problems discussed in this report.

## 1.2. Frame construction and maintenance

Once a decision is made concerning the target population, either the sample design must be based upon an available sampling frame(s) or the frame(s) must be constructed specifically for the study. Dalenius (1985) notes the following three important properties of a frame:

- Makes it possible to compute estimates concerning a population which is sufficiently "close" to the target population,
- Serves to yield a sample of elements which can be unambiguously identified, and
- Makes it possible to determine how the units in the frame are associated with the elements in the (sampled) population.

The first stage of sampling is usually dependent upon a frame consisting of a physical listing of units. This may be a list of names of individuals, establishments, institutions, counties, cities, streets, etc., or a list of numbers attached to city blocks, land area segments, houses, pages, or any number of other unique, definable entities. However, as Kish (1965, p. 53) notes, a "*Frame*" is a more general concept: it includes physical lists and also procedures that can account for all the sampling units without the physical effort of actually listing them." Deming (1960) cites one exception to the making of a list of sampling units, i.e., when a watch is used to sample time intervals during which customers leaving a store are interviewed.

The units listed in the initial frame may not correspond to the units about or from which information is sought. Often, additional frames are needed for successive stages of sampling in order to progress from available sampling units to the units to be contacted or measured. For example, areas may be selected from a listing or array of all blocks in an area frame. Housing units inside sampled areas may then be listed and sampled in order to achieve a listing of persons to be sampled that are members of the target population from which information is sought.

A more complex example is the procedure for selecting items to be priced in the Consumer Price Index. The sample of priced items is selected from items sold by a sample of outlets which, in turn, was selected from a list of outlets created from information provided by interviews with consumer units in addresses sampled from the decennial census, new construction permits, and area listings. In this case, interviews are conducted in a sample of housing units to create a sample frame of establishments, not a population frame, from which a sample is selected. Within the sampled outlets, probability methods are used to select increasingly more detailed classes of goods until a particular item is selected. A complete list of all the items available for sale is never constructed.

A variety of sampling frames utilized by agencies of the Federal Government is presented in table 1. Associated information related to construction and survey use of the frames is included.

In practice, with the exception of area frames consisting of land segments, the target population a sampling frame purports to represent is constantly changing. For a one-time survey, when it is desirable to obtain data for a specific point in time, this fluctuation is not usually critical, assuming the frame represents the near truth relative to the time of interest. It becomes more critical for ongoing surveys. While for such surveys, panel maintenance rules are inevitably applied so that the frames remain representative of the changing target populations, these rules are often difficult to apply comprehensively because of funding limitations and/or methodological complexities (see appendix A.1 and A.4). The result is that, over time, any panel may no longer be representative of the target population that is to be measured. Thus, resampling from a current frame usually occurs regularly for such surveys. A current frame is usually available because a procedure for updating the frame is formulated during the panel survey's design process and so is in place at the time the first sample is selected. This updating assures that the frame remains representative of the target population over time. For example, a universe file has been established for the Producer Price Index survey. The primary purpose of this file is to provide up-to-date establishment information including name, location, industry classification, employment, and other pertinent items. These data are obtained via telephone interviews during a frame refinement process or by personal visits to collect data only from sampled units (see appendix A.4, section III).

Not all frame maintenance procedures address problems of coverage. Some, such as the removal of inactive units, are geared toward sample efficiency. Still, neglecting such procedures can affect coverage. For example, deaths on a frame may be sampled but are not likely to respond. If they are treated as active units and data are imputed for them, bias is introduced. Therefore, it is proper to consider frame maintenance methods in more detail (see section 1.2.2). Before

Table 1: Selected Sampling Frames Used For Federal Surveys

FRAME NAME	PRIMARY FRAME SOURCE	DEFINED : COVERAGE : PERCENT	CREATION : DATE	SUPPLEMENTARY : FRAME SOURCE(S)	PERIODICITY : OF UPDATE	SURVEYS : SELECTED : FROM FRAME	FRAME : UNITS	VARIABLES : AVAILABLE FOR : STRATIFICATION
Structure TAR Structure PRELIST	: Decennial Census : of Housing	: 100 :	: 1980 :	: New construction : permits	: 10 years :	: CPS, CPP, CE, : SIPP, NCS	: Addresses :	: Geography, : demographics
ED Summary Tape & Block Summary Tape	: Decennial Census : of Housing	: 100 :	: 1980 :	: New construction : permits	: Annually :	: CPI-Housing :	: Blocks & : ED's	: ED summary : demographics, : geography
UI	: State unemployment : insurance filings	: :	: Annually :	: :	: :	: OSH, SCA, PATC, : CES, ECI, AWS, : OES, PPI, IWS	: Establish- : ments	: SIC, : employment, : geography
Oil & Gas Well Operators	: FEA lists, Federal & : State records	: 100 :	: 1978 :	: State & Federal : production re- : porting agencies, : industry press	: Annually :	: EIA-23 :	: Addresses :	: Production, : volume, : State
DOD Master Gain/ Loss File	: Lists submitted by : four active DOD : services	: 100 :	: 1971 :	: Service record : updates	: Quarterly :	: Personnel : surveys	: Persons :	: Service branch, : military unit, : geography, : demographics
National Master Facility Inventory	: State agencies, : professional : organizations, : NIMH	: 100 :	: 1962 :	: Census of inven- : tory, Agency : Reporting Systems	: 2-3 years :	: National : Nursing Home : Surveys, HDS, : NMES	: Establish- : ments	: Bedsize, : certification, : geography, type : of ownership
SSEL	: IRS 941 payroll file, : 1972 Economic Census, : Company Organization : Survey	: 100 :	: 1973 :	: SSA birth files, : IRS 941 payroll : file, current : economic surveys	: Quarterly :	: Retail, whole- : sale and ser- : vices, COS, : Economic Census, : CBP, Research & : Development, : shipments of : defense oriented: : industry	: Companies, : establish- : ments, EIN : number	: SIC, employment, : payroll, sales, : BMF activity, : geography

Table 1: Selected Sampling Frames Used For Federal Surveys (Continuation)

FRAME NAME	PRIMARY FRAME SOURCE	DEFINED COVERAGE PERCENT	CREATION DATE	SUPPLEMENTARY FRAME SOURCE(S)	PERIODICITY OF UPDATE	SURVEYS SELECTED FROM FRAME	FRAME UNITS	VARIABLES AVAILABLE FOR STRATIFICATION
Census of Manufactures	SSEL	100	1987	None	5 years	ASM, Plant capacity surveys, CIR, pollution abatement	Establishments	SIC, shipments, employment, geography
Census of Agriculture	Previous census, IRS 1040s, NASS lists	100	1987	List from trade associations, State & Federal agencies	5 years	Farm & ranch irrigation, horticulture census are examples	Farms	Geography, size (e.g., reported acreage)
Fraud Intelligence File	Customs entry documents	100	Monthly	New customs entry documents	Monthly	IPP	Import transactions for establishments	TSUSA's transaction dollar value, month of trade
Census Export File	Shippers export declarations	100	Monthly	New documents	Monthly	IPP	Export transactions for establishments	Schedule B number, transaction dollar value & month of trade
NASS Area Frame	Maps, photos and satellite images	100	1965	Local knowledge & detailed maps	20 years	Crop acres & production, livestock inventories, grain stocks inventories	Land areas	Land use
NASS List Frame	List of farm operators	50-80	1981	Surveys and administrative records	Annually	farm costs & returns, cost of production, farm labor, objective yield samples, farm numbers & land in farms	Names associated with agricultural operations	Agricultural variables by size



doing so, however, it is useful to note some additional distinguishing features of sampling frames.

Not all sampling frames are maintained over time, even those for ongoing surveys. In fact, the frames created for many sampling operations are discarded once samples are selected and approved. The sample that is representative of the frame at the time it is selected does not remain representative of the population of interest over time and neither does the frame from which it was drawn, if not maintained. When and if a new sample is selected, it is first necessary to construct a new frame that represents the current target population. An example is the use of the Census of Manufactures as a frame for the selection of the Annual Survey of Manufactures sample. The Census of Manufactures represents the manufacturing establishment population at a point in time and thus is not subject to change until the next Census of Manufactures, in 5 years. It serves as the primary, but not the exclusive, frame source for the Annual Survey of Manufactures, and is itself a derivative of the Standard Statistical Establishment List (SSEL). Once the sample for the Annual Survey of Manufactures is selected, it undergoes coverage updating each year, but no updating to the census frame can be done until the next Census of Manufactures. When the next census is completed, it will serve as the new frame for the next sample selection. The new census, while conceptually an update of the old census, is in fact developed from the latest version of the SSEL, which itself made use of the prior census results (see table 1).

For other sampling operations, the frames are evolutionary; that is to say, they are not fixed nor are they instantaneous creations. Instead, they evolve from periodic updates to a previous version of the frame. Each sample is taken to represent the target population at the reference time; however, the frame is maintained and updated to reflect the continuity of changes in the population it covers. In this context, frame maintenance is part of an iterative procedure, with results of a given survey contributing to changes in the frame from which subsequent cycles of samples are drawn. Two examples of this type of frame are the unemployment insurance (UI) file maintained by the Bureau of Labor Statistics (BLS) and the SSEL file maintained by the Bureau of the Census. Both files maintain a current profile (ownership, mailing address, Standard Industrial Classification (SIC) code, physical location, etc.) of economic entities in the United States. The UI file is updated quarterly using employer reports to the UI system. The UI file is supplemented with quarterly data for multiple reporting units, and the SIC and county codes are verified on a rotational basis, one-third of the establishment population each year. The SSEL is updated on a continuous basis using a variety of sources, including administrative records from the Internal Revenue Service (IRS) and Social Security Administration (SSA), the economic censuses conducted every 5 years, the Company Organization Survey conducted annually (except during a census year for most multiunit companies), and the many current economic surveys conducted by the Bureau of the Census. The UI file serves as a sampling frame for most BLS establishment surveys (see appendix A.4), while the SSEL is the underlying frame for most Bureau of the Census' economic censuses and surveys (see appendix A.1).

Other examples of evolutionary frames are two frames maintained by the Energy Information Administration (EIA). The Oil and Gas Well Operator List is used as the frame for the Annual Survey of Crude Oil, Natural Gas, and Natural Gas Liquids. A list of firms selling petroleum products is used as the frame for two surveys: The Annual Fuel Oil and Kerosene Sales Report and the Monthly Petroleum Products Price Report. Information to update these frames comes, in part, from responses given by operators and firms on their survey submissions. Information from several other sources, including the triennial Petroleum Product Sales Identification Survey, is also used in adding, deleting, or modifying entries on the appropriate frame.

One other point is worth noting. Files whose primary purpose may be to serve as a sampling frame may serve other functions as well. For example, UI data collected by the States is used primarily to administer the Federal-State UI system. Additionally, the file provides a base from

which to estimate the wage and salary component of national personal income and the gross national product. In addition to being a sampling frame, the SSEL serves many other purposes. It must fulfill the needs of many different survey programs with many different requirements. Because of this diversity, the amount of information included is limited, so the SSEL is not always used as the direct establishment frame source for sampling operations at the Bureau of the Census. For example, the Current Industrial Reports surveys are selected from a frame created from the Census of Manufactures. These surveys are commodity surveys, and, for the most part, the population of interest is all producers of the particular commodities covered by the survey. Primary producers of these commodities can be identified on the SSEL; i.e., those establishments classified in industries which include those commodities, but the SSEL does not contain information on quantity or value of those commodities. More importantly, secondary producers cannot be identified, e.g., a steel plant which also happens to produce leather shoes would not be identified as in the scope of a survey to estimate shoe production. For some surveys, the contribution of secondary producers could be significant. The Census of Manufactures, on the other hand, contains product data which allow all known producers to be identified, and for this reason sampling frames are created directly from it. The underlying basis for this census, of course, is the SSEL.

The remainder of section 1.2 contains a discussion of coverage errors associated with the creation and maintenance of physical lists as sampling frames like those included in table 1. Section 1.2.1 gives a classification of frame errors as put forward by Kish (1965) and modified by Lessler (1980). The problems of maintaining or updating a sampling frame to reflect changes in the covered population over time are addressed in section 1.2.2. The concerns and procedures discussed are also relevant to the creation of a physical list of population elements which is to serve as a sampling frame. Section 1.2.3 treats the special case of frame updating or creation by means of matching and merging multiple lists to create a single more current or complete frame.

#### 1.2.1. Classification of frame errors

Kish (1965) states a "frame is perfect if every element appears on the list separately, once, only once, and nothing else appears on the list," and classifies possible frame errors into four types: Missing elements, clusters of elements appearing on the list, blanks or foreign elements, and duplicate elements.

In a detailed presentation of errors associated with frames, Lessler (1980) classifies six types of error: The four types that Kish discusses, plus incorrect auxiliary information, and information insufficient to locate target elements. Incorrect auxiliary information can affect the coverage of the frame if the information is used to define subpopulations or subframes. Information insufficient to locate target elements does not reflect a coverage error in the frame, but may result in a coverage error as discussed under rules of association in section 2.1.1.

Missing elements. The omission of units in the target population causes greatest concern. Because units are missing, no examination of any sample from the frame will reveal the nature of the missing component of the population. Research conclusions may be erroneously extended beyond an incomplete frame on the frequently tenuous assumption that missing units are like or very similar to those on the frame. This assumption is to be distinguished from the assumption often made for sample estimation purposes that survey nonrespondents are like respondents. When this assumption about the frame used is not clearly revealed in research reports, the research community receives misinformation, as mentioned in section 1.1.

Missing units are most commonly the result of the following situations: Absence from sources used for frame construction, failure to report to an administrative system, births (new to relevant population), and zero units by definition. All of these circumstances might contribute to a conclusion that missing units are not like others which are included in the frame. Because it may

be extremely expensive to attempt to obtain complete coverage, an organization may or should show the missing component to be a trivial proportion of the total or institute some form of estimation procedure to account for the missing portion of the population. This is especially true when the missing units are suspected of being unlike the included units.

Examples of list frames considered very nearly complete for survey purposes include the UI file of business establishments, the SSEL, Oil and Gas Well Operators, the Department of Defense Master Gain/Loss File, and the National Master Facility Inventory (NMFI) for health services. (Refer to table 1 for some selected characteristics of these and other frames.) Some organizations such as the Department of Agriculture and the Bureau of the Census maintain area frames that are considered complete, since all areas of land in the United States are contained within the frames. Therefore, all activities occurring within the United States are theoretically reachable through these frames. However, completeness does not imply the frames or the surveys which use these frames are free of coverage errors.

Clusters of elements appearing on list. A frame is ideally composed of individual sampling units with known characteristics which identify or link to reporting units. The initial sampling units may be known to consist of clusters of subunits which can be incorporated into a sampling design. An example would be a listing of single-family dwellings that contains some duplexes. Another example is a list of farm operator names of which the vast majority represent a one-name/one-farm relationship but some represent a one-name/multiple-farm relationship. Jessen (1978) describes four different relationships between what he refers to as frame units and observation units. These various relationships introduce complexity into the survey process. There is a definite possibility for coverage error if field representatives have not been thoroughly trained in the proper procedures for handling clusters of reporting units associated with a single sampling unit.

Blanks or foreign elements. If a frame is created or an existing list modified for a particular one-time survey, elements on the list which are blank or are not members of the population of interest should be removed. However, most Federal surveys are repetitive or ongoing, and many frames are used for more than one survey. Thus, quite often it is appropriate to retain elements on the frame which previously were members of the population of interest for at least one survey. For a discussion of frame creation and maintenance procedures designed to deal with inactive or out-of-scope frame elements, see section 1.2.2.

Duplicate elements. Duplication of units on the frame may result in overcoverage, i.e., some members of the population are represented more than once. Population totals may then be overstated and means could be biased. Moreover, multiple representation of units on a sampling frame leads to sampling inefficiencies. There are, however, survey procedures which may be employed to identify and compensate for frame duplication (e.g., see Gurney and Gonzalez 1972).

Data collection may be complicated in the face of suspected frame duplication by the necessity of obtaining additional information in order to allow matching with the frame to find other frame elements representing the same population unit. For example, a farm name may be present on the list in addition to the farm operator's name. In the case of partnerships, any enterprise may have multiple representation through the names of individual partners. The necessity of obtaining these names and cross-checking against the frame lengthens the interview and complicates the survey process (see section 2.1.1, Location errors). The Producer Price Index Establishment Universe Maintenance System was developed for the Producer Price Index survey as a means of minimizing duplication as well as other sampling frame problems. It captures all changes made during frame refinement and collection feedback (see appendix A.4, sections III-V).

Undetected duplication resulting from nonsampling errors made during data collection or frame-check activities may result in a biased survey estimate. The extent of the bias depends upon the amount of duplication for which no adjustment is made and the size of the units involved. For example, a business enterprise may exist in the form of a vertically integrated company having a pyramid structure. Individual units may then maintain their own books on number of workers or value of production and contribute to the next higher unit in the structure. The parent unit may have the relevant data pertaining to the entire organization. The effect of not detecting the relationship among these sampling units depends upon which units happen to be included in the sample and how the structures of their operations compare to those of the remainder of the population.

Incorrect auxiliary information. Great care must be exercised when units are intentionally excluded from a frame because they are not thought to be members of the population of interest. Errors in frame variables, like size, type, class, or location of unit, could cause valid units to be excluded. For example, the SSEL file contains a relatively large number of records for which no industry classification has been assigned. These unclassified units become missing units on the various frames which are derived from the SSEL, since frame eligibility is first determined on the basis of industry classification. A major effort is made prior to each census year to code the unclassified units that have accumulated on the SSEL since the previous census, including, as a last resort, mailing an inquiry to an establishment to obtain a description of its activity. Little is done between census years because of the cost, but the business surveys which use the SSEL attempt, on a sample basis, a yearly classification, since experience has shown that most unclassified units are ultimately coded to their domain.

For a discussion of frame creation and maintenance procedures relevant to misclassified elements, see under section 1.2.2, Misclassified elements.

#### 1.2.2. Frame maintenance

In this section, frame maintenance procedures are discussed with reference to the kinds of coverage error described in the previous section. These procedures can be classified as follows:

- Adding new frame elements or births,
- Eliminating or identifying inactive frame elements or deaths,
- Correcting misclassified frame elements,
- Identifying existing frame elements no longer in scope, or in scope for the first time, and
- Determining whether or not elements have combined with other elements or have split from existing elements (e.g., change in ownership, mergers, and divestitures in an economic setting).

Each of these updating procedures is discussed in turn below. The discussions address the effects on the frame of failure to update. Distinctions are made between updating procedures intended to determine the current status of existing frame elements, and those intended to identify elements not previously known to exist. In addition, procedures that update the frame as a whole are distinguished from those that may update only a subset of the frame.

New frame elements. When the research population is dynamic, it is important that the frame which represents it be updated to reflect births. Samples drawn from frames which are not updated for births can result in serious biases, especially if simple weighted estimates are to be used (see discussion of missing elements in section 1.2.1).

One effective method for detecting new units is to canvass periodically the existing frame elements. As an example, the larger (50 or more employees) multiunit companies on the SSEL are canvassed on a yearly basis (with the exception of the census year) via the Company

**Organization Survey.** A proportion of the smaller companies is also canvassed in years other than either the census year or the year following. Companies are queried as to whether or not they have started new operations. However, companies do not always specify whether a newly listed establishment is a new entity (birth) or represents the purchase of an existing plant. If a plant is treated as a birth and sampled when, in fact, it had a chance of selection under another name or code, bias can result. (See appendix A.1 for additional details.)

A second method of identifying new units results from coverage maintenance operations performed for samples selected from the frame. This method, like the first, uses canvassing, but only of the sampled portion of the frame units. As part of the questionnaire administration process in nearly all surveys, inquiries are made about the status of the sampled units and whether any changes in their status have occurred since the last data collection period. Although the inquiries are targeted to sampled units believed not to be births, sometimes incidental information about other units (including births) can be obtained. This is obviously more a random than systematic approach for identifying new units. Inquiries made in the Annual Survey of Manufactures of single-unit companies provide an example of the use of this approach. Each sampled single-unit company is asked whether any additional plants operate at its location or whether the company owns any additional plants or is owned by someone else. The purpose of these inquiries is to determine whether or not the company is a multiunit company. If the single unit does identify other locations, these may well be establishments which are new or which were not previously known to exist.

Establishments are also added to the SSEL through new employer identification numbers (EIN's) received from the Internal Revenue Service. New numbers do not necessarily imply new establishments, however, as existing plants often request new numbers. The SSEL does not distinguish between the two. Duplication on the file of a plant under both a new and an old EIN will soon resolve itself, as the old EIN will eventually show no payroll data and will be dropped. Survey designers need to be able to identify the true births, however, and this requires additional work. In the Annual Survey of Manufactures, for example, classification cards are mailed to all manufacturing-coded establishments given a new EIN in an attempt to determine whether the establishments are births or existing plants. Only a sample of true births is added to the survey panel (see appendix A.1, section II).

Administrative records are also used to add establishments to the UI file. New business establishments are required to file with the State employment security agencies. However, there is a time lag between filing and being added to the UI file. Units added to the UI file are not necessarily births. Mergers, changes in ownership, branch offices, etc., may sometimes be assigned new UI account numbers. In an effort to address this problem, State agencies are trying to identify units which are legal predecessors and successors within the UI system. In addition, units which do not meet the legal UI requirements, but are still essentially the same economic units, may be identified as predecessor/successor by the States. In the meantime, the Producer Price Index survey annually uses an automated process whereby the new incoming UI file is compared to the universe file. If an establishment fails to match a unit on the universe file, it is added to the universe file with a special code (see appendix A.4, section III).

The Bureau of Labor Statistics (Grzesiak and Tupek 1987) has conducted several studies of business births in conjunction with its Current Employment Statistics program. The usefulness of the UI file as a sampling frame for new businesses is constrained by the delay between the time a business first hires employees and the time it enters the UI file. A study of all 12,983 UI accounts (the sampling frame for this program) assigned by Florida for 3 months in 1984 found almost 80 percent were new accounts without predecessors. The study focused on determining the length of time between a business's first liability for UI coverage and its entrance on the UI file, which depended on how the State discovered the employer and whether the employer had a predecessor. The median lag-time for all new accounts was found to be 120 days. A study was

also conducted in New York to develop a methodology for identifying new businesses using the UI system and to construct new procedures for estimating the employment of new businesses for incorporation into the Current Employment Survey. The median lag-time in New York was also found to be 120 days, with 93 percent of the establishments having fewer than 10 employees.

Record checks with outside sources can also be used to identify birth elements. Generally, these checks do not allow one to distinguish between new establishments and previously missed establishments, but in either event they provide information for updating the file, and reveal elements that had no chance of inclusion in previously-selected survey samples. One example of an outside source frequently used in record checks of establishment frames is the trade association list.

Several methods have been used for identifying birth elements to the National Master Facility Inventory. Each method has relied on State agencies' lists of facilities. (See appendix A.3, section II for details.)

The traditional method for including births in housing unit surveys is to update field-generated listings of sampling or reporting units within sampled geographic areas. Initial listings are usually made just prior to the first interviewing period and are subsequently updated through a recanvass to correct errors and to add newly constructed units.

The Bureau of the Census uses this approach in some geographic areas for the housing unit surveys it conducts. However, in most areas, births are included by sampling building permits. (See appendix A.7, section II for details.) Sampling building permits results in a significantly lower sampling variance, since large housing projects can result in very large clusters of sampling units being added during the alternative field-listing update process.

However, the building permit files do not identify illegal new construction, conversions, and new mobile home placements; nor do they identify new special places, such as dormitories, fraternity houses, boarding houses, and public housing. To illustrate, it was estimated for the 1985 Annual Housing Survey that approximately 25 percent of all new mobile homes were missed (Schwanz 1988a). In the Survey of Income and Program Participation, the undercoverage of new mobile homes in the building permit file was estimated to result in a 1-percent underestimate of the number of households in poverty (Singh 1989). In 1976, the undercoverage of births in the building permit frame was estimated to be about 2.3 percent (U.S. Department of Commerce 1978a). Since the Bureau of the Census normally uses this building permit frame for sample augmentation over a 10-year period, e.g., 1975-85, undercoverage may increase substantially over this time span. (For a more complete description of the procedures used by the Bureau of the Census to identify and sample dwelling units created after the last decennial census, see appendix A.7.)

Another methodology for capturing new housing construction and reducing undercoverage in sampled geographic areas is the half-open interval procedure. Instead of listing all units within the sampled area, a string of  $k$  units is listed in a predetermined order. The string begins with a designated unit from the original frame and is bounded by the  $k$ -th unit that was reported in the original frame (Montie and MacKenzie 1978). A modification of this procedure was used in the 1977 National Health Care Expenditure Survey and the 1980 National Medical Cost Utilization and Expenditure Survey. Cursory analysis indicates the approach may be limited in its ability to capture new construction (Adams 1989).

Inactive frame elements. Efficient sampling dictates that inactive units or deaths on a sampling frame be identified. In the initial construction of a frame, deaths in the various sources used to construct the frame should be identified and, if needed, removed. Existing frames should be updated periodically to remove or flag units that are no longer active. Failure to identify deaths

on a sampling frame does not necessarily result in overcoverage, but, as was noted earlier, biased sample estimates can result if an inactive element is sampled and imputed for when no response is obtained.

It may be desirable to retain inactive units on a frame for a certain period of time because of estimation considerations or because it is desirable to have a history of elements available. When doing so, the inactive units should be identified either through flagging or partitioning to a distinct death subfile. In the Producer Price Index survey frame, a death is defined as either a sampled unit which has been identified as out of business or out of scope, or any existing frame unit that remains unmatched when the next UI file is compared to the universe file. All deaths are removed from the universe file and added to the death file (see appendix A.4, sections III and V).

Two methods, both of which were mentioned in connection with identifying births, are particularly effective in identifying whether units are still active or in existence. The periodic canvass of existing frame elements is one. The frequency of these updates varies. For the COS, a canvass of the multiunit portion of the SSEL is conducted annually (except for the census year) for the large companies (50 or more employees). The smaller companies are periodically canvassed as well, but not on a yearly basis and not all in the same year. The companies are asked to identify any listed plants that are now out of operation. The National Agricultural Statistics Service's list frame is canvassed within 5 years of the preceding canvass. The UI file is updated quarterly using a census of UI reports. Units not paying taxes for some period of time (4-8 quarters) are removed from the frame. On the other hand, fixed frame elements are not canvassed at all, since frames such as the Census of Manufactures are not updated. It should be noted, however, that the companies in the Census are canvassed to update the SSEL.

Inactive elements are also identified through maintenance operations performed on samples drawn from the frame. In nearly all survey panels, sampled units which are no longer active are routinely identified through inquiry. Maintenance operations performed on samples are more likely to reveal changes in status of known elements (from active to inactive) than to reveal births (new elements), although this method can be used for both purposes. The information obtained using this method can then be used to update the source file (frame). The SSEL is continually being updated from information obtained by the many Current Business Reports surveys, the Current Industrial Reports surveys, the Annual Survey of Manufactures, and many other economic surveys. The UI and National Agricultural Statistics Service list files are updated in this fashion as well. It is important to note that this method yields information for updating only the sampled units and for adding new elements revealed through survey operations. It does not permit updating the entire frame.

Misclassified elements. A problem with many frames is not that elements are missing, but that they are misclassified or are not classified at all with respect to one or more variables. This assumes importance if the variable or variables that are misclassified determine either the elements eligible for sampling or the subpopulations for which estimates are produced. Housing occupancy status (vacant or occupied), geographic codes, SIC code, age, race, and gender are examples of such variables.

For economic surveys, where the population distribution of the variable to be estimated is usually extremely skewed, some measure of size is often used in the sample design as either a stratification variable or for sampling with probability proportional to size. Incorrect values for the variable(s) used to derive an establishment's measure of size can result in its being placed in the wrong stratum and in other sampling inefficiencies. To illustrate, in the Annual Survey of Manufactures, which has a probability-proportional-to-size design, an arbitrary certainty stratum is defined to consist of all frame establishments with total employment greater than 250. Since estimates are not published for this stratum, the erroneous inclusion or exclusion of an



establishment in this stratum because of an incorrect employment size value does not bias a survey estimate, but the resulting sampling error may be different than expected.

Coverage error at the estimation stage due to misclassification will often be confined to sublevels. For example, in economic surveys like the Annual Survey of Manufactures (ASM) with independent sampling across SIC's, an inaccurate SIC code could result in undercoverage if the code identifies the unit as nonmanufacturing. However, if the SIC code is incorrect, but within manufacturing, only industry level estimates will be affected (see appendix A.1, section II).

Incorrect classifications may result from errors in data input during frame creation, but more often they occur because changes in frame units are never detected by the surveying organization. To update industry and location changes on the UI file, BLS conducts a refiling survey in which SIC and county codes are verified for each unit of the universe on a 3-year rotational cycle.

In the 1977 Economic Censuses, misclassification studies were conducted for both the employer and nonemployer segments of the administrative records frame. For the employer segment, a subsample of 5,505 out-of-scope employer cases on the 1977 Economic Censuses Master Sample were mailed the Economic Census General Schedule to complete. An estimated 3.1 percent of out-of-scope employer establishments with 0.4 percent of the employees and 0.3 percent of the annual payroll were found to be misclassified as out of scope (Hanczaryk and Sullivan 1980). An estimated 12 percent of the nonemployer establishments were misclassified as out of scope, resulting in a 20-percent underestimate of nonemployer receipts.

A majority of the employer misclassifications resulted from errors in the SIC on the administrative file. In some industries, an establishment may be comprised of distinct but related activities, e.g., construction and real estate. However, it is classified into an industry on the basis of which activity yields the highest percentage of total receipts. For example, an establishment with 45 percent of receipts attributable to construction and 55 percent to real estate sales would be classified as in the real estate industry and excluded from the Economic Censuses. If the percentages had changed between the time of coding and the census, the establishment would be misclassified, since construction was within the scope of the census.

The evaluation study found that many of the establishments misclassified on the basis of out-of-scope SIC codes had in-scope Principal Industrial Activity codes on the administrative file. Therefore, a significant drop in the misclassification rate could have been achieved through the use of another variable. On the other hand, missing or incorrect tax return Principal Industrial Activity codes were responsible for a majority of the nonemployer misclassifications. Again, in this situation, the use of additional tax return information could have substantially reduced the misclassification rate.

While many of the procedures discussed previously can, to a degree, aid in the identification of misclassified elements, none of them is really intended to address this problem comprehensively. Their effectiveness depends on the specific variable of interest. For example, although the SIC code is part of the information collected in the Company Organization Survey (COS), it is not likely that the companies would routinely verify that the right code is assigned. Nor can the Bureau of the Census determine the validity of the code, since the COS does not collect detailed data.

Another procedure to handle misclassification is used in the Producer Price Index survey. Since some establishments can easily modify their capital equipment to produce a different product, depending on demand, the 4-digit industry classification can change. Thus, industries for which a high proportion of this type of misclassification occurs are treated as related SIC's and sampled



at the same time. The sampled establishments are then assigned the proper SIC (see appendix A.4, sections III-V).

An auxiliary variable that is used as a measure of size can sometimes cause coverage error if it is incorrect. For example, in the Producer Price Index, the employment value on the UI file is used as the measure of size. Some establishments are reported with zero employees. In order to ensure that all units have a positive probability of selection within the probability-proportional-to-size sample design, two employees are added to the employment value of each unit. For EIA's survey of active oil and gas well operators, companies on the frame having no known production are sampled, so that they are represented along with those operators having known current production.

Unclassified elements are unique in that it is known at the outset that they exist. Knowing this information and perhaps the reason(s) for lack of classification allows the surveying organization to design a strategy for obtaining codes. This strategy may not necessarily be useful for resolving the lack of classification on all frames. For example, most of the units on the SSEL without SIC codes do not have them because the SSA is unable to assign them a code when their applications for new EIN's are received. Prior to a census year, the Bureau of the Census makes a concerted effort to code these records. This includes identifying key words in an establishment's name which might identify its activity and, when this fails, culminates in the mailing of a classification card to the establishment which asks for a description of its activity. A certain number of records remain uncoded. Although new EIN's are obtained by the Bureau on a continuing basis, little is done for them between census years because of the cost. However, attempts are made to classify a sample of them for the business surveys because most unclassified units fall within their population of interest.

Out-of-scope elements. Closely related to the problem of misclassification is the problem of out-of-scope elements, i.e., elements that if properly classified would not be part of the population of interest. They differ from the type considered above in that if they were properly classified, they would be dropped from the frame. Out-of-scope cases generally arise because historically they were coded in error. It may be, however, that their status has changed so that they are no longer part of the population of interest (see appendix A.2, section II). As with death elements, the presence of out-of-scope elements on a sampling frame does not result in any biased sample results should they be sampled (assuming the survey processing identifies them as out of scope), but it does compromise the efficiency of the sample.

Split-out or combined frame elements. The composition of elements constituting a frame will often change over time. This is especially true for establishment frames, where, for example, individual plants are bought and sold by companies, two or more companies merge, or companies divest. But it is also true for frames of housing units and households. In these instances, frame maintenance properly includes activities which update or modify the frame to account for compositional changes.

Compositional changes do not necessarily affect the number of units on establishment frames and, thus, the overall coverage of the frames. Indeed, it is likely that no changes in plant activity occur at all. If both the sampling unit and the reporting unit are the establishment, it is really not vital that the corporate owner of the plant be known as far as data collection is concerned. From a coverage point of view, however, ownership may be important because the sample status of a sold establishment often depends upon the status of the buying company. Also, in some economic surveys, establishment records are combined into company records for sampling purposes. Thus, there are a variety of other reasons, some coverage related, which mandate maintenance of proper identification of plant ownership.

For the SSEL, the annual canvass of the multiunit establishments by the COS is the prime basis for maintaining the identification. The COS provides a list for each company of all the known establishments of the company and requests that the company verify and update the list by indicating any new establishments it has opened, the name and seller of any additional plants it has acquired, the name and purchaser of any plants it has sold, and any plants it has closed down. COS processing identifies the new owners (successors) of sold plants and the old owners (predecessors) of bought plants and corrects the records for those companies as well. Similar activities are conducted for virtually every census/sample survey enumeration at whatever level the reporting is done.

The UI file is another example of a broad-based file which is supplemented with quarterly data for multiple reporting units. The frame units that have undergone changes in composition are identified as new owners with predecessor relationships and old owners with successor relationships. Thus, for establishment-based surveys, the company is queried about changes in the operation of each establishment, including whether or not it has been sold or leased to any other company.

In the Producer Price Index survey, four economic characteristics are used to define a unit. Pertinent establishment data are obtained via telephone interviews during the frame refinement process. Any change in the composition of an establishment is captured on the universe file. If an establishment is split, the new portion is treated as a birth and is added to the universe file with a special code. If a unit is sampled and during the first interview a portion is identified as either split or sold, it is treated as a field-created sampling unit and data are collected (see appendix A.4, sections III and IV).

Agricultural frames generally do not have this kind of multiunit situation. Each operation (farm or ranch) is defined by one common land operating arrangement. This lowest common denominator precludes the necessity of keeping track of elements within farm units. However, bookkeeping arrangements covering farms in a hierarchical management system may necessitate periodic monitoring to be sure each unit is accounted for but not duplicated. A more complete description of counting rules for agricultural surveys is provided in the Quarterly Agricultural Surveys case study (appendix A.5).

Likewise, for the Census of Population and Housing frame, the identification of basic addresses does not change at the transfer of ownership from one person to another. Other identification problems may exist, such as how many separate living quarters really exist at a particular location.

### 1.2.3. Match-merging of independent source lists

In many of the examples of updating procedures discussed above, it was noted that outside source lists or files were used to update a primary frame. Among the problems arising from the use of such lists not addressed above are those associated with matching and merging such lists to update primary frames. These problems deserve special mention because they affect frame completeness.

The two general classes of error that can occur when combining lists are: Erroneously adding an element already in the frame and erroneously removing a qualified element from the frame. The two types of error are not equally problematic, since a more stringent set of rules can govern the deactivation of a frame element than governs the incorporation of a new element.

The updating process entails some formal matching between the primary frame and the source list. Various identifiers may be utilized in the match operation (Fellegi and Sunter 1969, Scheuren and Oh 1985). These identifiers may have different degrees of precision ranging from

very precise (e.g., EIN) to less precise (e.g., name, address), and it may be that successive matches are attempted on each level of identifiers. At the end of this process, records on the primary frame and on the update list can be allocated into three mutually exclusive parts:

- a. Records which are classified as matching (i.e., appear on both frame and source list),
- b. Records on the primary frame that do not match to the source list, and
- c. Records on the source list that do not match to the frame.

Some records which match may represent false matches. Depending on the quality of the source list, i.e., whether it is truly free of duplicates and out-of-scope units, false matches can lead to failure to add a unit or, more rarely, failure to identify a potential death.

Depending on the conceptual and operational fit between the source list and the in-scope population, the failure to match some frame records to the source list may or may not be a problem. The ascribed completeness, timeliness, and accuracy of the source list are all important in deciding whether unmatched entities have died and should be eliminated from the frame (or flagged) or to leave them on the frame and let sampled deaths be revealed during data collection. If any sampled entity is a death, that fact will become apparent when it cannot be found, although in a mail-out/mail-back survey, any such unit may be presumed to be a nonrespondent for which data are imputed.

Source list records not matched to the frame represent potential additions to it. Although these records can just be added to the frame, it would be more prudent to try to determine whether they are duplicates of existing units or are out of scope.

Many variants of the process described above occur in survey situations. At one extreme, there is no attempt to determine if firms with newly issued EIN's are already on the SSEL under an older EIN. At the other extreme is the procedure followed for the National Master Facility Inventory of inpatient health facilities, in which names and addresses of facilities from source lists that do not match the primary frame are automatically added to the frame.

Problems related to the timeliness of the source lists can arise. Since the population of interest is not static and events are cumulative, combining untimely information from source lists with that on an existing frame can lead to numerous errors. Ideally, one wants the frame resulting from the match-merge to include all elements in the population of interest and to exclude elements which are not. For example, if a survey is to be conducted of firms currently in business, one does not want to rely on a historical file of all businesses that does not denote firms no longer in business. To do so would be to risk including these firms in the sample. Less error prone, but still problematic, is the use of source lists containing information for units that existed at any single point during the year. Units that exist throughout the year will be included, along with those born during the year. But, those that died will also be included. Should such a list be used to update a frame for a sample survey in the succeeding year, units no longer in existence may be included. In such instances, the reasons units no longer exist may be extremely useful pieces of information, if they can be obtained. For example, has a firm simply gone out of business, did its name change, or was it purchased by another company? Otherwise, when deaths are sampled and revealed as deaths during data collection, frame records for these deaths can be flagged as inactive or deleted from the frame, as appropriate. Also, the use of less than timely source lists can result in the addition of unknown out-of-scope units that will remain on the frame to plague subsequent surveys.

### 1.3. Sample design strategies to minimize coverage error

In the previous section, the discussion focused on coverage error associated with sampling frames. Solutions to problems arising from the limitations of available frame sources are a major

challenge to the survey design statistician. Colledge (1989) identifies and discusses 26 specific coverage and classification problems faced by Statistics Canada in its Business Survey Redesign Project, as well as possible alternative solutions.

This section presents some sample design and estimation options available to survey designers in dealing with recognized deficiencies in a frame. The options discussed are: Defining the target population to equal the frame population, random-digit dialing sampling, multiple frame sampling, sampling rare populations, and estimation procedures.

#### 1.3.1. Defining target population to equal frame population

While it is important not to imply coverage of a wider population than the one covered by the frame(s), it is more important to make concerted efforts to reach every member of the original target population, even if this means using additional frames or more expensive procedures. Only intolerable expense or practical impossibilities should be grounds for narrowing the defined target population, as discussed in section 1.1.

Hansen, Hurwitz, and Jabine (1963) provide an example of how a coverage problem for a survey about truck ownership and operation was handled. When it became clear that State motor vehicle registration records did not include all trucks being operated and that coverage of truck registration varied by State, the scope of the study was redefined as registered trucks instead of all trucks.

#### 1.3.2. Random-digit dialing sampling

One household sampling method used to avoid omission of households with telephones is random-digit dialing (RDD) (Waksberg 1978). The use of telephone directories as sampling frames often results in unacceptable levels of undercoverage because they omit unlisted numbers for some nontypical portions of the population. With RDD, a sample of telephone households is located through the use of randomly generated telephone numbers. In this way, only those households without telephones are omitted. For many surveys, this could be considered a trivial exclusion. In others, differences between telephone and nontelephone households may have a profound effect on the characteristics being measured. For example, measures of poverty and income from entitlement programs would most likely be biased because households in poverty or receiving such income are less likely than other households to have telephones. The collective experiences of numerous researchers and survey statisticians who have used RDD are presented in Groves, et al. (1988).

An extensive discussion of the health characteristics of persons in telephone and nontelephone households is presented by Thornberry and Massey (1988). Data from the National Health Interview Survey indicate that those in the nontelephone U.S. population are more likely to suffer disability days, chronic conditions, and hospitalizations than those in the telephone population. At the same time, those without telephones have fewer visits to physicians and dentists and are much less likely to have private health insurance. These findings are consistent with expectations, given there are disproportionately more low income families in the nontelephone population. The authors note that for most characteristics, the differences between the values for the telephone households and the total population are small because 93 percent of all households can be reached by telephone via RDD. However, estimates for certain population subgroups could be severely biased when based on an RDD survey. The authors note that an RDD survey seeking information on preschool aged children would exclude about 12 percent of them, and also almost one-third of such children living in poverty.

An example of favorable results from using RDD is reported by Williams and Chakrabarty (1983) for the State of Michigan portion of the 1980 National Fishing, Hunting, and Wildlife

Associated Recreation Survey. Parallel surveys were conducted utilizing an RDD sample and a subsample from previous Current Population Survey samples which did not depend upon presence of a telephone. The report points out "the socioeconomic characteristics and the sportsmen variables between the two studies do not reflect any substantially important differences." However, there were differences in results for "nonconsumptive users," i.e., wildlife-related activities outside of hunting or fishing. These activities were highly related to the geographic location of the user, so findings may result from the geographically restricted nature of the expired CPS samples compared to the unrestricted nature of the RDD sample.

A study by McGowan (1982) on telephone ownership in the National Crime Survey sample contains evidence that the exclusion of nontelephone households has a significant effect on the measurement of crime victimization in the United States. In this instance, the use of RDD without a supplemental frame to provide a sample of nontelephone households would be unacceptable.

### 1.3.3. Multiple frame sampling

Coverage may be improved through the use of multiple frames. Sometimes, no single frame fully covers the target population and merging independent source lists would be impractical. In this case, separate probability samples from different frames can be used to expand coverage beyond any available single frame. (Additional frames may also be used to increase sampling efficiency if coverage is already sufficient.) The use of multiple frames entails two assumptions (Hartley 1962):

- Every unit in the population of interest belongs to at least one of the frames, and
- It is possible to record for each sampled unit whether or not it belongs to the other frame(s).

The first assumption requires linkage between the sampling frame units and the target population. Application of rules of association to accomplish this linkage is needed when sampling from any frame (Hansen, Hurwitz, and Jabine 1963). When multiple frames are used, sampling units are often different between frames. This is of no consequence as long as the different sampling units lead to a common reporting unit during the survey. Complete coverage of the reporting units should be equivalent to complete coverage of the population of interest. The rules of association, from sampling units selected to reporting units tabulated, must ensure the representation of each population element once and only once in the final estimates. Field representatives must be equipped with clearly defined rules that can be communicated to respondents to achieve this unique representation.

Difficulty with the first assumption, given a need to use multiple frames, arises because concurrent application of different rules of association may be required of field representatives depending upon which frame supplied the sampled unit. Potential errors in associating sampling units with reporting units are discussed in section 2.1.

The second assumption requires that frame membership be known for each population unit. Nonoverlapping frames are a special case, wherein each population unit is assumed to be on one and only one frame. The statistical theory for this special case is essentially the same as for stratified sample designs.

In the case of nonoverlapping frames, each frame represents a different, unique segment of the total population. The principal consideration here is that the same reporting unit not be included in more than one of the sampling frames. In this way, estimates for the nonoverlapping frames are additive to the greater population of interest. Examples of the use of nonoverlapping frames within government are the Current Employment Statistics survey as well as other establishment surveys conducted by the Bureau of Labor Statistics. The primary frame for employees comes

from UI reports filed with State employment security agencies. The UI frame covers about 98 percent of wage and salary employment in the United States. Supplemental, nonoverlapping coverage comes from the Interstate Commerce Commission for interstate railroad employees. Another example is found in the National Cancer Institute's epidemiology studies, where the under-age-65 group is selected from a frame of driver's license records and the over-65 group is selected from a frame of Medicare records.

Most housing unit surveys conducted by the Bureau of the Census, e.g., American Housing Survey, Current Population Survey, Consumer Expenditure Survey, National Crime Survey, and the Survey of Income and Program Participation, use a combination of frames (U.S. Department of Commerce 1978a). In areas where building permits are required and maintained by a local government and the Census of Population and Housing addresses contain street names and numbers, the census lists are used as the basic sampling frame. A sample of building permits is also selected to cover housing units built after the census. Conceptually, these two frames are nonoverlapping even though they refer to the same land areas. In other areas of the country where permits are not available for sampling or the census address lists are considered inadequate, land areas are sampled, and an address list is created by field representatives. For a discussion of coverage errors in this listing process, see section 2.2.1.

Use of overlapping frames in the application of multiple-frame survey methodology mandates that extraordinary attention be paid to potential errors in the survey process. A population (reporting) unit may fall within any or all frames utilized. Sampling units are selected from each frame and linked by rules of association with corresponding reporting units. Each reporting unit must ultimately be represented exactly once across all frames utilized. This may be accomplished either directly through a matching process to remove duplicates or indirectly by weighting adjustments. (The latter tends to be far less costly.) Duplication because of multiple representation or omission through failure to account for the unit in at least one of the frames can result in serious coverage errors.

Sampling from overlapping frames is most commonly done when an area frame and an overlapping list frame are available. The area frame is generally designed to provide complete coverage by including as sampling units all land parcels which encompass the population of interest. The list frame is nearly always incomplete, a common attribute of lists, but its use provides certain sampling efficiencies which enable the multiple frame survey to provide the same precision at a much lower cost than would an area frame survey alone.

Examples of the area/list dual-frame survey approach may be found in the Department of Agriculture (in nearly all inventory and economic probability surveys conducted by the National Agricultural Statistics Service) and in the Bureau of the Census (in the Monthly Retail Trade Survey and the Services Annual Survey). The Department of Agriculture's application of this approach for the Quarterly Agricultural Surveys of crops, hogs, and grain stocks inventories is a typical illustration of the linkage requirements with multiple frame sampling (see appendix A.5).

An important special case occurs when an existing complete frame is used in conjunction with a list of telephone numbers. This general case has been discussed extensively in the literature. See, for example, Lepkowski and Groves (1986) and Biemer (1983). Important special cases are considered by Lepkowski (1988).

#### 1.3.4. Sampling rare populations

There are two known procedures to compensate for undercoverage that are especially useful for surveys of rare or elusive populations: Network sampling and capture-recapture methodology. Both are briefly described below.

Network sampling used in conjunction with multiplicity estimation (Sirken 1970 and Sirken and Levy 1974) relies on a known set of relationships (or links) between members of the population. Network sampling, unlike more traditional sampling, uses links which extend beyond the usual sampling or reporting unit by building rules for more extensive sampling. One example of such extended rules is the sibling rule, where sampled members are asked not only about themselves, but also about all brothers and sisters not living in the same household.

In network sampling, a sample is drawn from an established frame using a probability sampling procedure. The sample is then contacted and interviewed to determine which sampled members have the characteristic of interest. Sampled members are then asked about the set of related individuals having the characteristic being studied. In this way, several members of the population are covered in one interview. Since this procedure is potentially prone to increased response error or item nonresponse, the names and addresses for the related individuals who are said to have the characteristic of interest are often obtained, so that the individuals can be contacted directly. This technique is best known for its potential to improve efficiency when the characteristic of interest is rare. However, it also has the potential for improving coverage when people are reluctant or unable to provide information about themselves and when the sampling frame is incomplete (Sirken 1983).

A survey to collect data on recent decedents is an example of a population unable to provide information about itself. The traditional methodology would be to collect information at the household that had been the decedent's place of residence. A network sampling approach would be to collect information at the household of a surviving spouse, siblings, or children residing in the county of the decedent, either instead of or in addition to at the decedent's last place of residence. Sirken (1983) reports on the results of experiments conducted in North Carolina to compare coverage between network sampling and traditional sampling (Sirken and Royston 1976). The traditional method missed 29 percent of the deaths; reports from decedents' relatives' households alone missed 22 percent of the deaths; and reports from both decedents' former residences and decedents' relatives' households missed 15 percent of the deaths. Emigrants are another group of people for whom network sampling can improve coverage because they cannot report for themselves.

Network sampling can be useful to improve coverage on incomplete sampling frames (Sirken 1983). Persons with no fixed address would usually be missed by traditional sampling but could be identified by relatives or friends. Also, if institutions or Armed Forces barracks are not included in the sampling frames, network sampling can be used to find persons living in these otherwise uncovered places.

Use of network sampling requires that the number of population units eligible to report each sampled individual be known. This number is used in the estimation process to adjust the probability sampling weight for each sampled unit.

Also known as dual system (or multiple system) estimation, capture-recapture methodology assumes that one or several frames have less than perfect coverage of the population, and that the amount of undercoverage is unknown. Capture-recapture methodology is essentially a counting technique and is used to determine the number of individuals in a population, or the number of individuals with a specific characteristic in a known population.

The population to be studied is defined independently of any frames, but at least two overlapping frames are needed to make an estimate of the population size. Membership on any frame is modeled as a stochastic event and, for two frames, membership is also assumed to be an independent event between frames. The two frames are matched, or a sample from one frame is matched to the entirety of the other frame. An estimate is then made using the number of persons estimated to be in the first frame ( $N_1$ ), the number of persons in the second frame ( $N_2$ ),

and the number found in the match to be in both frames ( $M$ ). The estimator (Marks, Seltzer, and Krotki 1974, p. 15) of the population size ( $T$ ) is  $t = N_A N_B / M$ .

A number of assumptions are required to satisfy the model which generates this estimator; some of the assumptions can be relaxed if more lists are available for sampling. Lists can include administrative records, but the model requires the assumption that membership in the records system is a random event, an assumption that usually does not hold. (References include: Casady, Nathan, and Sirken (1985); Czaja, Snowden, and Casady (1986); and Cowan, Breakey, and Fischer (1988).)

For a general treatment of strategies for sampling rare populations, see Kalton and Anderson (1986). See also appendix A.2, section III.

### 1.3.5. Estimation procedures

Estimation procedures which compensate for known coverage error in frames may be used to decrease the bias of survey estimates. Improving frame coverage is always better than using these estimation procedures. One such procedure is ratio estimation or benchmarking; another approach is multiple frame estimation.

The Bureau of Labor Statistics employs a benchmarking procedure to revise monthly employment estimates from the Current Employment Statistics survey (U.S. Bureau of Labor Statistics 1989). Sample estimates are compared each year with later summarizations of mandatory UI reports filed by employers. The UI data, which serve as a benchmark, are an aggregation from the same source as the microdata used to construct the frame from which the sample was selected, except that the benchmark data are one year newer. Hence, the benchmark file takes into account new firms or changes in industrial classification to ensure more accurate coverage. The completeness of the UI administrative data affords the opportunity to analyze and adjust for frame deficiencies (Thomas 1986).

Most of the current surveys conducted by the Bureau of the Census use ratio estimation to projected population totals by age, sex, and race. For further discussion of the procedure as applied to the Current Population Survey, see appendix A.7, section VI.

The use of multiple overlapping frames requires the use of an estimator which may be written as follows for the two-frame case, where frame sizes are known but overlap domain size is unknown (Hartley 1962):

$$Y = (N_A/n_A)(y_A + py'_{AB}) + (N_B/n_B)(y_B + qy''_{AB}),$$

where subscripts  $A$  and  $B$  denote the two sampling frames,  $N$  and  $n$  are the population and sampled units, and  $Y$  is the total of some variable to be estimated. Subscripted  $y$ 's are estimated totals from the two frames ( $y_A$  based on units uniquely in frame A,  $y_B$  for units only in frame B, and  $y'_{AB}$  the estimated total for units in both frames as measured by the frame A sample; while  $y''_{AB}$  applies to units common to both frames from the frame B sample), and  $p$  and  $q$  are weights which sum to one. In this way, the estimates unique to each frame are added to a weighted combination of those units common to both frames. The parameter  $p$  is selected so variance is minimized subject to a cost function reflecting differences in sampling from each frame.

A common application of this estimator utilizes a complete area frame A and an incomplete but more efficient list frame B to generate a screening multiple frame estimate of the form:

$$Y = (N_A/n_A)y_A + (N_B/n_B)y''_{AB}$$



where the estimate for the units unique to the area frame (nonoverlapping domain) is added to the list frame estimate for the units common to both frames. In this case, the parameter  $p$ , from the general formula above, is zero, and  $q$  equals one. Other terms disappear because no units exist on the list that are not contained within the area frame.

It is easy to see in the simple form of the multiple frame estimator the importance of properly determining whether or not a unit is represented by one or both frames. Unrecognized overlap between the frames produces duplication in the estimate, while improper designation of a unit as overlapping results in omission.

#### 1.4. Evaluation methods

One method of measuring the degree of frame coverage error is comparative analysis. Comparative analysis can occur at two levels. The first is a macro-level evaluation, which compares known population values with totals derived from summing characteristics for each sampling frame unit. The second type of analysis is performed at the micro or individual sampling unit level. This most often involves matching of data available from different sources for individual units.

##### 1.4.1. Macro-level analysis

How do totals associated with sampling units compare with other measures of the target population? Suppose we have an area frame. The sum of the areas in individual sampling units or segments should match closely with the measured area of the total frame, e.g., county, State, or other target area. The National Agricultural Statistics Service electronically digitizes clusters of area sampling units and verifies that the accumulated total is within 0.5 percent of the published land area for each State (Cotter and Nealon 1987).

Tortora (1987) notes that with two frames, one a complete area frame, a process quality control evaluation of a list frame is possible through the use of survey data. For example, list coverage of the number of farms or land in farms can be estimated by the sizes of the overlap and nonoverlap domains from the area frame. Likewise, the number of out-of-scope list units can be estimated from the samples in each frame. Monitored over time, the measures of list frame performance will provide knowledge and control of list coverage.

Similarly, the number of names in a list frame can sometimes be compared with census counts for the population of interest. Generally, the information available on every sampling unit is very limited and only gross comparisons with known population totals can be made. More often, totals estimated from sample surveys can be compared to similar quantities from other sources in order to provide measurements for the frame. Two of the most common sources are census and administrative files.

Reconciliations are made between economic census totals and corresponding totals from the Current Industrial Reports annual survey for census years. Similarly, the Department of Agriculture conducts a continuous survey program for the agricultural sector and routinely compares inventory and production estimates with those obtained in the agricultural censuses conducted at 5-year intervals by the Bureau of the Census.

The Bureau of the Census utilizes still another macro-level approach for frame completeness evaluation called demographic analysis. With this method, demographic data from various sources are used to develop expected values for the population as a whole and by race, age, and sex to compare with the census counts. This procedure relies on aggregate statistics of birth, death, immigration, emigration, past censuses, Medicare enrollment, and other sources to provide estimates of net census coverage errors for broad categories at State and national levels

(Fay, Passel, and Robinson 1988). The estimate of the net undercount of the legally resident population in the 1980 decennial census is 1.0 percent using this procedure (p. 26).

The mandatory UI reports filed by employees with their State employment security agencies are the primary source of information for the BLS Universe File. This file is used both for sampling frame maintenance and during the estimation process. For example, comprehensive totals from the Universe File at the SIC or SIC/size class level can be used to evaluate the sampling frame inadequacies caused by lack of timeliness for the Current Employment Statistics Survey. Births of new firms (economic units which have begun operations since the time of frame construction) and inaccuracies at detailed levels resulting from changes in SIC codes contribute to differences between the survey frame and the target population. The degree of undercoverage during the time lag until discovery of new units depends upon the number and size of operations entering the target population. During the estimation process, an updated Universe File is used to ratio adjust the estimates; the reference period for the updated Universe File is one year later than for the Universe File used as a sampling frame. Evaluations of survey data versus target population totals show that only minor revisions apply to Current Employment Statistics Survey results (Thomas 1986).

Several studies have been made of business births and job generation which indicate the importance of measuring employment in new businesses. Roughly 800,000 businesses are formed each year, creating 2,500,000 new jobs. While jobs in new businesses constitute a small fraction of total nonagricultural payroll employment (annual average employment of 104,300,000 in 1988), they are a substantial portion of net new jobs (2,800,000 from 1986 to 1987 and 3,200,000 from 1987 to 1988). An analysis of Dun and Bradstreet credit rating information (Birch 1979) showed that small businesses (20 or fewer employees) accounted for two-thirds of net new jobs between 1969 and 1976. Other studies using Small Business Administration files at the national level or files at the State level have shown that more than half of the net employment growth came from small businesses or business births (Armington and Odle 1981; Teitz, Glasmeier, and Svensson 1981; Connor, Heeringa, and Jackson 1985). These studies show the importance of including new businesses in establishment surveys of employment.

#### 1.4.2. Micro-level analysis

Micro-level analysis of sampling frame units implies direct matching or linkage of the same units found in more than one source. Given a common reference unit, be it person, housing unit, or business, the information available from an administrative file, a census, or survey source should verify and enhance the data associated with the unit.

The U.S. Department of Commerce's "Report on Statistical Uses of Administrative Records" (1980) includes four case studies of projects utilizing comparative analysis between surveys, census data, and administrative records. All four of these studies utilize matching between files at the individual record level to assess coverage problems and illustrate the kinds of sampling unit evaluations possible across frame sources.

Such record-matching studies are performed for statistical purposes only. In general, strict laws govern the release of unit-level data collected for statistical purposes. For example, Title 13, U.S. Code prohibits the Bureau of the Census from releasing information that allows the identification of survey or census respondents. The same law allows the Bureau of the Census to obtain administrative information from other agencies and organizations in support of its statistical activities, including computer matching. When the statistical and administrative records are matched, the resulting files are afforded the same guarantee against disclosure as the statistical records.

Two of the studies are concerned with the use of multiple administrative record systems to serve as frames for integrated samples and to provide basic information for record-check studies. The first is the Linked Administrative Statistical Sample Project involving samples from the records systems of the Internal Revenue Service, the National Center for Health Statistics, and the Social Security Administration. In the second project, "The Use of Administrative Records in the Survey of Income and Program Participation (SIPP)," the Department of Health and Human Services and the Bureau of the Census use administrative files from the Aid to Families with Dependent Children program, Supplemental Security Income recipients, and the Basic Educational Opportunity Grants applicants.

The third study describes how the Bureau of the Census uses the Internal Revenue Service tax files for persons aged 17 to 64 years and the Medicare file for persons 65 or over in conjunction with the Current Population Survey sample to provide reliable estimates of coverage error in the 1980 decennial census for certain geographic areas and some socioeconomic categories. The final study, "Record Linkage in the Nonhousehold Sources Program," is a large-scale record check by the Bureau of the Census which matches census returns against administrative records for drivers' licenses from State departments of motor vehicles and against registers of resident aliens supplied by the Immigration and Naturalization Service to reduce census undercoverage.

A comparison of macro- and micro-level analysis procedures for evaluating coverage of the 1980 Census of Population and Housing and results of their application can be found in Fay, et al. (1988).

## CHAPTER 2

### COVERAGE ERRORS OCCURRING AFTER INITIAL SAMPLE SELECTION

Errors occurring after the initial selection of a sample from a frame can be classified broadly into three types: Incorrect association of sampling with reporting unit(s), listing errors, and other nonsampling errors.

In the first section of this chapter, the discussion focuses on coverage errors which occur in the process of associating a unit selected to be in the sample with a unit from or about which data are to be collected. Misidentification of the physical unit selected for sample and misclassification of the sampled unit as either a member or nonmember of the eligible population are two types of such errors. Some of the coverage problems inherent in the survey process, which by necessity cannot be instantaneous, are discussed under the rubric of temporal errors.

At some point in the sampling process, a frame or master list of subunits within the last selected unit may not exist, so field data collection agents are asked to compile such a list for further sampling or interviewing. The more stages of listing and sampling required to move from the initial frame to the unit for which a response is desired, the greater the opportunity for mistakes. This is particularly true when sampling control moves from a centralized location to individual field representatives. Listing errors are the topic of section 2.2.

In section 2.3, the discussion focuses on the coverage implications of other types of nonsampling errors, such as recording errors, volunteer responses from nonsampled units, and failure to elicit a response for a sampled population unit (nonresponse).

#### 2.1. Incorrect association of sampling with reporting unit(s)

Sampling units used in the frame construction process often differ from the reporting units which are or which provide data about members of the target population. Relating the sampling unit to the reference (reporting) unit for respondents in a concise and unambiguous manner is an important challenge for the survey design statistician. The ability of the field representative and respondent to avoid coverage errors at this stage of the survey process depends upon the clarity with which sampling and reporting units of interest can be defined in survey materials and field representatives' instructions.

The concern here is with the rules which must be applied by field representatives to determine what reporting unit is specified by the sampling unit. Fuzzy rules that do not consistently lead the field representative from a single sampling unit, selected with known probability, to one or more unique reporting units while retaining the correct known probability of inclusion, can result in coverage errors. When the rules of association are very complex, coverage errors are more likely to occur. Similarly, survey concepts that are unclear or objectionable to respondents can also impair coverage.

Rules of association delineate the relationship between sampling units and the final reporting units. For the purpose of this report, rule-of-association errors have been classified into three basic types: Location errors, classification errors, and temporal errors.

##### 2.1.1. Location errors

Location errors result from the difficulty of associating sampling units with reporting units when the sampling units are not uniquely or clearly defined or when they are difficult to locate. In a housing unit survey with a good listing of addresses, it is extremely difficult to determine if field representatives have conducted interviews at wrong addresses. Error studies of sampling unit

listings for the CPS have shown the overall quality to be quite good. A very small percentage of errors has been attributed to such things as defective address units (0.31 percent), errors in records that caused serious problems for field representatives (0.09 percent), and combinations of two separate addresses into one (0.10 percent) (U.S. Department of Commerce 1978a).

In surveys of establishments or firms, the task of associating sampling units with reporting units is more difficult. For instance, if a part of a larger organizational entity is sampled, the response may reflect the entire firm rather than the sampled part. Sometimes, as has been noted in the Annual Survey of Manufactures, respondents will request forms for additional (unsampled) establishments within their company, then proceed to complete and submit those forms. The same survey has also determined that respondents may combine data from several plants, some of which are not included in the survey, and some of which are individually selected. In an attempt to avoid this problem with one segment of the population in the Annual Survey of Manufactures, single-unit companies are asked whether they have operations at any additional locations. If so, they are asked to list those locations and report minimal data, such as kind of business, sales, payroll, and employment. They are also asked whether data for these units have been included as part of their response. If so, the data are allocated to each of the sampling units during data reduction. The additional locations are researched to determine whether they had any previous chance of selection or had been omitted from the frame.

In overlapping multiple frame surveys where the opportunity for multiple representation within and between frames is high, appropriate coverage of the population is heavily dependent upon rules of association. These rules must define how to disaggregate data among sampling units in such a way that the reporting unit is represented only once. For example, the reporting unit of interest in most Department of Agriculture surveys is land operated by farmers and ranchers. Farm operators provide the required survey information regarding inventory of crops and livestock for the acres they operate. The list and area frames for these surveys utilize two different sampling units. Parcels of land sampled from the area frame must be linked with the operators farming the land. Names of potential operators sampled from the list frame must be verified as operators and the amount of land operated must be determined (Beller 1979 and appendix A.5).

There are then several sources of coverage error which must be addressed by rules for associating sampling units with reporting units within and between the frames used in the Department of Agriculture surveys. Within sampled land segments from the area frame, field representatives must identify the appropriate operator or operators in order for State office personnel to check whether or not that land is also represented in the list frame. For units sampled from the list frame, the proper linkage between sampled name and land operated is necessary to ensure the same land would be identified as overlapping if found in the area sample. Finally, the names of operators and any multiple names associated with land parcels (e.g., farm business names or partners farming together) must be obtained in order to determine which sampled units are common to both frames.

The method of handling duplication within the list frame can influence the rules for matching elements between frames. Each of the names associated with a sampled parcel of land in the area frame must be matched against the list frame. There are then at least three alternative procedures for assigning response data to the overlapping and nonoverlapping domains.

- Prorate the data to each name associated with the reporting unit so the proportions allocated to the overlap and nonoverlapping domains correspond to the proportions of names on the list and not on the list, respectively (partial overlap procedure).
- Associate all data with the list frame so long as any name associated with an area tract of land is found on the list (list-dominant procedure).

- Accept an area parcel as belonging to the list domain only if all names appear together as a single list sampling unit (area-dominant procedure).

### 2.1.2. Classification errors

In many surveys, the population of interest is a subset of the sample frame. The decision to define a frame unit as within the scope of a survey and to conduct an interview may be made in the office on the basis of administrative data on the sampling frame, by a respondent in a mail survey, or by a field representative who has a set of rules to determine if a sampled unit should be interviewed. The set of rules may take the form of procedures or a screening questionnaire. To the extent frame units or sampled units are misclassified as out of the scope of the survey, undercoverage occurs.

Misrepresentation of out-of-scope units as members of the target population results in overcoverage. This can normally be detected during data collection when a response is obtained. If changes in the sampling units have occurred between construction of the frame and the first interview, the first interview must include procedures to determine whether the unit should be interviewed. If a nonresponding unit is treated as a missing unit through imputation or weighting when it is not a member of the target population, then overcoverage results. In this case, deaths of sampling units present a special difficulty during the time they remain sample nonrespondents.

Some examples of surveys which collect data from only a subset of the sample are given in table 2.

Table 2. Scope of frame versus population of interest for selected surveys

<u>Survey(s)</u>	<u>Frame</u>	<u>Scope of Frame</u>	<u>Population of Interest</u>
CPI Housing	Area segments	All housing units	Civilian noninstitutional housing units, excluding military and public housing units
CPS, NHIS, CE	Decennial census addresses	All year-round housing units	Occupied housing units and special places, excluding vacant units
IPP	Fraud Intelligence	All imports of value >\$1,000	Imports of domestic companies, excluding foreign governments and transshipments
OES	UI, railroads	All employers	Only 1/3 of SIC's sampled each year, excluding private households in agriculture
R & D	SSEL	All SIC's	All manufacturing SIC's, excluding selected nonmanufacturing SIC's
NMCUES	State Medicaid rolls in 4 states	All Medicaid beneficiaries in 4 states	Civilian noninstitutionalized Medicaid households in 4 states

Misclassification in many economic surveys results from inaccurate SIC coding on the administrative files used as sampling frames. Misclassification in most housing unit surveys results from field representatives classifying an occupied housing unit as vacant.

The estimation of misclassification rates is an important tool for evaluating the effects of misclassification on coverage. However, misclassification rates are more often estimated for censuses than surveys. Therefore, the following discussion of misclassification--its magnitude, causes, and effects--is based primarily upon census evaluation projects.

In the 1982 Census of Agriculture, an estimated 3.1 percent of the estimated total number of farms in the United States were missed because they were misclassified as nonfarms, and approximately 3.7 percent were overcounted due to misclassification of nonfarms as farms (U.S. Bureau of the Census 1984). The farms which were missed because of misclassification accounted for approximately 23 percent of all estimated missed farms. Eighty-four percent of the misclassified undercounted farms had less than \$10,000 in sales. Misclassification in the Census of Agriculture may result from the respondent misinterpreting the instructions and/or census definitions. For example, some respondents feel their operation is "too small" or "for home use only" and should not be classified as a farm. Others think that because they have no sales of crops or livestock in the reference year or were only operating a farm for a few weeks in the reference year, they should not respond. Incomplete reporting by respondents and errors in census processing also lead to classification error.

In two separate evaluations of the 1970 Census of Population and Housing, 11.4 percent and 16.5 percent of the units initially enumerated as vacant were misclassified and should have been enumerated as occupied (U.S. Bureau of the Census 1973). The occupied-enumerated as vacant misclassification rate was approximately twice as high in multiunit structures as in single-unit structures. As a result of the National Vacancy Check study, 0.5 percent of the total population was imputed to adjust for the misclassification of occupied units as vacant.

As a result of the 1970 census misclassification problems, a followup of vacant and deleted housing units was conducted as a coverage improvement procedure in 1980 (U.S. Bureau of the Census 1987). This procedure found that an estimated 17.3 percent of the deleted units should have been enumerated, 7.5 percent as occupied and 9.8 percent as vacant, and an estimated 11.2 percent of the vacant units were misclassified, 9.1 percent as a result of enumerator error and 2.2 percent as a result of procedural errors. Enumerator errors occurred when enumerators visited housing units occupied before and during the census but classified them as vacant. Procedural errors occurred when housing units had been vacant early in the census-taking period, but were occupied before the end of the census by people who had not been enumerated at previous addresses. As a result of this study, 1,724,087 persons (0.76 percent of the total population) were added to the 1980 decennial census count.

In October 1966, the CPS reinterview survey concentrated on measuring the coverage errors made by field representatives (U.S. Bureau of the Census 1968). In the CPS, noninterviews are classified as type A if the housing unit is occupied by persons eligible for interview; as type B if the unit is either unoccupied but could become occupied, or occupied solely by persons not eligible for interview; or as type C if the unit is ineligible for the sample. In table 3, a cross-classification of the 325 reinterviews conducted during that time period indicates approximately 10 percent of the units originally classified as vacant should have been classified as occupied.

Table 3. Reinterview classification of units originally classified as noninterview: October 1966

<u>Original Classification</u>	<u>Reinterview Classification (percent)</u>		
	<u>Interview or Type A</u>	<u>Type B</u>	<u>Type C</u>
Type A	99.04	0.96	-
Type B	10.11	88.83	1.06
Type C	12.12	-	87.88

As a comparison, estimates of the reinterview classification of the 2,499 units originally classified as noninterviews during the period April to September 1966, when coverage was not the focus of reinterview, are shown in table 4. In this case, only about 3 percent of the units originally classified as vacant were found to be occupied.

Table 4. Reinterview classification of units originally classified as noninterview: April to September 1966

<u>Original Classification</u>	<u>Reinterview Classification (percent)</u>		
	<u>Interview or Type A</u>	<u>Type B</u>	<u>Type C</u>
Type A	96.57	3.26	0.17
Type B	3.45	95.87	0.68
Type C	12.12	-	87.88

The misclassification error rates for type B noninterviews were found to be higher when the purpose of the reinterview was specifically to examine coverage errors than when the reinterview was not specifically targeting coverage errors. Therefore, the CPS misclassification error rates presented in table 5, based upon the reinterview of 4,940 units in 1987, should be interpreted with this fact in mind (Newbrough 1988). In addition, the reader is cautioned to note that the relative standard errors for these estimates are high. Standard errors are not available for the 1966 data, but the standard errors for the 1987 percentages are given in parentheses below.

Table 5. Reinterview classification of units originally classified as noninterview: 1987

<u>Original Classification</u>	<u>Reinterview Classification (percent)</u>		
	<u>Interview or Type A</u>	<u>Type B</u>	<u>Type C</u>
Type A		0.12 (0.05)	0.06 (0.03)
Type B	0.43 (0.09)		0.67 (0.12)
Type C	0.02 (0.02)	0.12 (0.05)	

Beginning in 1985, the type B noninterview rate for the Current Population Survey and the Survey of Income and Program Participation began to increase gradually. Once identified, this became a concern, since most of the units were recorded as vacant. Table 6 provides the type B rates for the Survey of Income and Program Participation and the Current Population Survey for 3 years.

Table 6. Type B rates for the Survey of Income and Program Participation and the Current Population Survey, 1985-87 (percent)

<u>Survey</u>	<u>1985</u>	<u>1986</u>	<u>1987</u>
SIPP	14.35	14.43	15.20
CPS	14.85	15.29	15.74

Two potential reasons for the increase in type B rates were investigated. First, the increase may have resulted from the misclassification of type A noninterviews (refusals from occupied housing units) as type B noninterviews by field representatives. Second, the vacancy rate among housing units may have truly increased from 1985 to 1987. If the first hypothesis were true, misclassification of eligible units could be an important source of survey undercoverage. If the second were true, undercoverage would not occur. The investigation revealed nothing to rule out a real increase in the vacancy rate among residential housing units as the cause for the increase in type B noninterview rates (King 1988).

In many surveys, screening for eligible respondents takes place at sampled housing units. Volunteer responses may occur if the field representative mistakenly includes families or



individual respondents who are not eligible. An error like this may occur when the field representative fails to screen at all or misinterprets the screening requirements as the result of inadequate training. A good example of this type of problem is described in appendix A.2, section III (see also Anderson, Schoenberg, and Haerer 1988).

### 2.1.3. Temporal errors

Time inherently complicates the complete and accurate representation of any population of interest on a frame. There are many phenomena which affect survey coverage that are magnified by the passage of time. For example, different farm operators are sometimes listed at different times for the same farm. This could simply result from an error in respondents' understanding of the farm operator concept from one frame development to the next, or it could actually reflect a change in farm ownership between the time of one frame creation/update and the next. Rules of association to account for inevitable changes in the population of interest over time are necessary.

In housing unit surveys conducted by the Bureau of the Census, there are problems associated with the use of a separate housing permit frame (see section 1.3.3 for a description of multiple frame sampling and appendix A.7). One problem concerns units built around the time of the decennial census. Those units deemed to be completed at the time the 1980 census (about April or May 1980) were included in the census frame. New construction units are selected from building permits. A unit for which a permit was issued in November 1979 may or may not have been completed at the time of the census. A special study was undertaken to determine the average time elapsed between permit issuance and construction completion for different categories of housing and by region. The results were used to determine the starting dates for sampling permits, i.e., the point in time when the number of housing units with no chance of selection was approximately equal to the number of housing units with two chances of selection. Thus, there would be little net bias arising from selecting new construction units authorized before the census, if characteristics of units with no chance and two chances of selection were the same (see Statt, et. al. 1981).

The length of time involved in establishing a frame, selecting a sample, and fielding a survey inherently affects the survey's coverage, since changes in the population of interest occur from the time of inclusion on the frame to the time of interview(s). Temporal error results when the frame or sample is not updated to represent the population of interest for the survey's reference period. Time-related coverage errors due to field representatives and respondents should be examined from both cross-sectional and longitudinal perspectives. Cross-sectional coverage error results from unaccounted for changes in the sample population from the time of frame establishment to the first interview. Longitudinal coverage error results from unaccounted for changes in the sample population from the first to subsequent interviews. For both situations, resulting coverage error will supplement the coverage error resulting from any inadequacies of the initial frame (for examples, see appendix A.2 and A.3).

The length of reference periods, or the periods for which information is collected, also affect coverage. These periods may be defined as some particular time before frame establishment, as the time of interview, or as any time between the two. Since changes in the population of interest occur continuously, there will be discrepancies in the population interviewed and the population of interest which are dependent on the reference period. Reference periods which differ from the time of frame establishment inherently cause coverage problems. Field representatives and respondents contribute to the discrepancies as a result of confusion in interpreting reference period rules, carelessness in applying reference rules, and memory loss. For example, suppose the frame for a housing unit survey is established in January and the households' reference persons interviewed in March are asked to provide household composition for January. The field representative may not convey the reference period correctly so that the

respondent provides March household composition, or the respondent may not accurately recall the household composition in January. In general, using relatively short reference periods should improve respondent recall.

Even though a frame may be representative of the population of interest at a point in time and the sample representative of the frame from which it is selected, it is imperative that the sample for recurring surveys be updated to remain representative of the true target population across time. Coverage errors can occur if the types of procedures discussed in section 1.2.2 on frame maintenance are not followed.

In longitudinal surveys, special rules of association are necessary to account for changes over time. For example, there may be changes in certain attributes of units, so that a sampled unit is a member of the target population in one or more reference periods, but not in other reference periods. Seasonal businesses or items, for example, may be eligible for the survey in reference period A but not in reference period B.

It is fundamental in the design of recurring and longitudinal surveys to anticipate and prepare for births, deaths, mergers, splits, movers, etc., so that a representative sample over time is maintained. Such changes may be accounted for by adjusting interviewing or by adjusting the sample through sample maintenance or weighting. Additional resources may be needed to consistently and correctly update the sample in some surveys. For example, in the 1979 Income Survey Development Program (predecessor of the Survey of Income and Program Participation), it was found that just to follow movers who settled within 50 miles of a sampled primary sampling unit cost 6.66 percent of the total time charged and 10.25 percent of the total mileage charged. In the current Survey of Income and Program Participation design, movers are followed if they move within 100 miles of a sampled PSU or can be contacted by phone (King, Petroni, and Singh 1987).

The National Agricultural Statistics Service employs a frozen domain approach (Bosecker 1984) for its multiple frame surveys. The status of each sampling unit is determined for a base survey reference point (June 1). List and area frames are, in effect, frozen as they existed on June 1 for all following surveys until the next June. All existing farmland has a known probability of selection in one or both frames at the time of the base survey. Very nearly all new farm operations are born out of existing operations, so the National Agricultural Statistics Service allows the inclusions of births after the reference date through the death of an existing sampled unit. Original selection probabilities for the old unit apply to the new unit.

In many surveys, it is the responsibility of the respondent to notify the survey sponsor of changes. This is especially true in mail-out/mail-back surveys. As one example, a major coverage problem occurred in a Department of Energy survey of gas sold as a result of the time lag between sample selection and data collection. Some time after the sample was selected, some companies set up operating subsidiaries to sell their gas. Thus, when the parent companies reported zero gas volumes, undercoverage occurred. The coverage problem has now been rectified by appropriate procedural changes.

Respondents may not inform field representatives of changes either because concepts are not understood or respondent burden would increase to an undesirable level. Similarly, field representatives are charged with the responsibility of identifying, tracking, and updating the sample according to specified procedures if such changes occur. Field representatives may be unable to uphold this responsibility if survey concepts are not clear or training is inadequate. Also, field representatives themselves may opt not to update for a known change in status if a significant increase in workload would result.

The sample design also affects the ease with which status can be updated. For example, it should be easier for a field representative to locate a mover, obtain information about births, deaths, splits, and transfers if an area sample design has been adopted rather than a random-digit dialing system because neighbors or subsequent occupants of the sampled unit may be able to provide the field representative with information. A partial solution is to provide preprinted cards to respondents to be mailed to the central office upon change in status. This, however, assumes the cooperation and ability of the respondent to acquiesce in such cases (U.S. Office of Management and Budget 1986).

## 2.2. Listing errors

Surveys may or may not have the benefit of sampling from a frame containing the final sampling units. The two tables below give a selected representation of each situation in the Federal Government.

Table 7. Selected surveys in which the frame sampling unit and the final sampling unit are the same

<u>Survey</u>	<u>Frame Sampling Unit</u>	<u>Final Sampling Unit</u>
Annual Survey of Manufactures	Establishment	Establishment
Monthly Survey of Industrial	Company	Company
Natural Gas Delivery		
Annual Retail Trade Survey	Company or EIN (smaller units)	Company or EIN (smaller units)
Producer Price Index Survey	Establishment	Establishment

Table 8. Selected surveys in which the frame sampling unit and the final sampling unit differ

<u>Survey</u>	<u>Frame Sampling Unit</u>	<u>Final Sampling Unit</u>
Consumer Expenditure Survey	Housing unit (list)	Consumer unit
Current Population Survey	Land segment (area)	Consumer unit
	Address (list)	Housing unit
Quarterly Agricultural Surveys	Land segment (area)	Housing unit
	Operator name (list)	Land operated
Survey of Neurological Disorders	Land segment (area)	Land operated
	Institutions	Patient or guardian
National Hospital Discharge Survey	Hospitals (list)	Discharges

Two challenges to the designers of the surveys in table 7 are to ensure the unit selected is in fact the unit found and interviewed by the field representative, and the respondent reports for only the particular unit selected. Intensive reinterview procedures are usually required to detect errors in meeting these challenges (see section 2.1.3).

Surveys with a target population different from the first-stage field sampling units, like those shown in table 8, obviously face additional difficulties. The more stages involved in the sampling process to proceed from the initial field sampling units through intermediate listings to reach a final sampling unit, the greater the opportunity for coverage errors.

The following table shows for selected Federal surveys the types of units within sampling units which require field listing.

Table 9. Examples of surveys requiring field listing

<u>Survey</u>	<u>Type of Sampling Unit 1</u>	<u>Type of Sampling Unit 2 Listed Within Unit 1</u>	<u>Type of Sampling Unit 3 Listed Within Unit 2</u>
Annual Retail Trade Survey	Area segment defined by counties	Area segment defined by zip codes, roads, etc.	Retail service establishments
Consumer Expenditure Survey	Housing unit (list) Land segment (area)	Consumer units Housing units	----- Consumer units
Current Population Survey	Address (list) Land segment (area)	Housing units Housing units	Persons Persons
Epilepsy Survey	Area segment	Pharmacies	Prescriptions
Head and Spinal Cord Injuries Survey	Area segment	Hospitals	Discharge diagnosis
Quarterly Agricultural Surveys	Area segment	Land by operator	
White-Collar Pay Survey	Company	Establishments	Occupations

As can be seen by an examination of the table above, some surveys have several stages of field listing. Within area or land segments, listings may be made of residential dwelling units or land controlled by farm operators. Within establishments, lists of products, employees, or occupations may be created. Within addresses, lists of housing units, consumer units, or persons may be created. Whatever the type of unit being listed, the potential for listing error exists.

The types of error include failure to find units which should be listed, failure to classify a unit as being within the scope of the list, listing a unit which is not within the scope of the list, or listing a unit more than once. The causes for error include inaccurate source materials such as maps, incomplete or inaccurate instructions for listing, insufficient training, or failure to follow instructions.

#### 2.2.1. Area segment listing errors

Studies measuring error. The few housing unit survey studies which have been conducted to measure the error in various area segment listing processes reveal that the level of error is relatively small. For example, over the last 10 years, the annual gross and net error rates in listings of area segments for the CPS, as estimated from the ongoing field representative quality control program, range from 1.58 percent (s.d. 0.06 percent) to 4.01 percent (s.d. 0.10 percent) and 1.70 percent (s.d. 0.10 percent) to -0.50 percent (s.d. 0.06 percent), respectively (Schreiner 1987). The median difference between the number of dwelling units listed in sampled area segments for the National Nielsen Television Index Survey and the 1980 decennial census housing unit counts was found to be -0.4, while the median difference without regard to sign was 4.9 (Hawkes 1985).

Even though the gross error level may not be very high, an examination of the frequency and types of error indicates there are areas where improvement is warranted. The following table shows the distribution of errors found in the Nielson study (Hawkes 1985), which was conducted in 1982, 27 months after the 1980 decennial census.

Table 10. Comparison of A. C. Nielsen 1982 field canvass of housing units with 1980 census housing unit counts by block group or enumeration district (National Nielsen Television Index Survey segments only)

<u>Percent Differences</u>	<u>Number of Survey Segments</u>
Total number of segments	2,001
+30 or more	74
+20 to +30	40
+10 to +20	95
+5.5 to +10	121
+2.5 to +5.5	157
+0.5 to +2.5	195
-0.5 to +0.5	181
-2.5 to -0.5	271
-5.5 to -2.5	258
-10 to -5.5	257
-20 to -10	198
-20 to -30	86
Less than -30	68

In another study comparing listed units for the CPI Housing Survey to the 1980 decennial census, Jacobs (1986) found that BLS field agents listed approximately 3 percent fewer residential dwelling units than in the 1980 decennial census. Part of the difference, an estimated 2 percent, was due to a difference in scope, i.e., the CPI Housing Survey does not include public, institutional, and military housing. In the CPI Housing Survey study, members of the Washington, DC office staff independently relisted several hundred segments. A line-by-line comparison of 192 segments, the first returned for processing and, therefore, a nonprobability sample heavily skewed toward smaller and more rural areas, indicated that although the total number of units did not vary substantially, the number of units determined to be in common between the two lists was less than 90 percent. The major problem was found to be unnumbered units, where the lack of a house number and variations in written descriptions precluded a match between listings. The next most prevalent problem occurred in multiunit structures where the exact number of units in a particular structure varied between the two listers. Boundary determination was also a problem, but much smaller in magnitude than those just mentioned. The primary causes of most of the errors were inadequate procedures, instructions, and training. Some procedures, such as a windshield screening of very large segments, proved to be very error prone. An emphasis on procedures and training in map reading, boundary determination, and listing for unnumbered units and units in multiunit structures was recommended for future listing processes.

Statistics Canada conducted a quality measurement study of its area segment listing process for its Labor Force Survey (LFS) (Joncas 1985). A dependent check of initial lists of 46,857 residential dwellings in over 10,000 segments, which excluded those selected from apartment or special area frames, resulted in an estimated average segment quality of 98.5 percent. That is, 1.5 percent of the units which should have been listed were determined to be missing from the original list. As another measure of quality, 72.8 percent of the segments were estimated to be errorless in the LFS study. The percent of errorless segments varied from 42.4 percent for segments located outside block-faced areas of large cities or rural areas to 80.2 percent for self-representing urban areas or small self-representing urban towns where the segments correspond roughly to blocks, combinations of blocks, or block faces. The distribution of error types found in this study was as follows:

Table 11. Number of listing errors found in Labor Force Survey study (Statistics Canada)

<u>Type of Error</u>	<u>Number</u>
Total	1,044
Boundary misinterpretation (not listed, inside boundary)	71
Portion of road network missed	99
Dwelling thought to be excluded	156
Dwelling located in business structure	23
Missed or hidden dwelling	224
Boundary misinterpretation (listed, outside boundary)	298
Dwelling converted to business	28
Dwelling incorrectly included	92
Error not classified	53

An intensive coverage check of the Current Population Survey in 1966-67 found that most of the listing errors in area segments resulted from the incorrect determination of segment boundaries. "In some cases this was due to the interviewer's inability to read maps; in other cases, the map was worn and illegible or the boundaries no longer existed. Other important reasons for differences were failure to update a segment, failure to list units under construction, failure to canvass back roads and impassable roads, and segment was difficult to list." (U.S. Bureau of the Census 1968, p. 37.) The distribution of reasons for errors in the October 1966 portion of this study was as follows:

Table 12. Reasons units were added and deleted during reinterview, as determined by reconciliation--area segments only: October 1966

<u>Type of Error</u>	<u>Number</u>
Total	110
Boundary incorrectly determined (not listed, inside boundary)	38
Bad map (not listed)	16
Concealed	4
Appearance was deceiving	4
Miscellaneous addition	10
Error not classified (addition)	16
Boundary incorrectly determined (listed, outside boundary)	11
Definition of housing unit	3
List not updated at proper time	4
Miscellaneous deletion	1
Error not classified (deletion)	3

In the few listing studies for economic surveys which have been conducted, the level of error is somewhat higher. For example, the gross listing error rate for listable (retail, wholesale, and manufacturing) establishments within area segments is estimated at 4.0 percent annually (Konschnik 1987).

The NASS area frame for U.S. agriculture utilizes aerial photographs to delineate accurately land parcels using identifiable boundaries. Estimates of the number of farms are dependent upon identifying every resident in sampled segments who meets the farm definition (sales or potential sales of \$1,000 or more in agricultural products). Finding and listing these individuals in high-density housing areas is especially difficult and expensive considering the rarity of farm

operators in residential neighborhoods. To reduce cost, NASS utilizes a procedure called the skip technique to screen residential segments (typically 0.1 square mile in size). Some houses are accounted for indirectly by asking interviewed persons whether any of their neighbors is engaged in any agricultural activity.

Special studies were conducted in 1986 and 1987 to determine the error rate associated with the skip technique (Matthews 1988). A procedure was instituted whereby each randomly selected house in a subsample received an interview. Estimates for this procedure and the skip technique were then compared. Undercoverage for the number of farms using the skip technique was estimated at 5.4 percent in 1986 and 6.3 percent in 1987. (Commodity estimates are not dependent upon the estimate for number of farms and so are little affected by operators found living away from their agricultural land.)

Budget considerations have prevented using this subsampling procedure. However, a multiplicity sampling procedure based on land operated, excluding land for residential purposes, was under investigation by NASS in 1989 to make listings unnecessary in residential areas (Bosecker and Clark 1988). Farm estimates would be based on an estimator utilizing acreage devoted to agricultural purposes rather than where the operator lives.

In addition to estimating the overall level of undercoverage due to area segment listing errors, two studies have identified some specific biasing effects. Hawkes (1985) noted that the Nielsen listing process resulted in a lower proportion of one-person households than the decennial census. For the Retail Trade Survey, the ratios of net error in sales caused by field representatives listing sales of establishments that should not have been listed and vice versa to area sample sales estimates and total sales estimates are estimated as -5.0 percent and -0.3 percent, respectively, for the 12-month period ending October 1986.

In the CPS, a housing unit is defined as "a house, an apartment, a group of rooms, or a single room occupied as a separate living quarters, or if vacant, intended for occupancy as a separate living quarters." The CPS makes extensive use of address listings that were compiled for enumeration districts (ED's) during the previous decennial census. Some other ED's are canvassed just prior to the initial CPS interviews. In either case, the field representative must apply the housing unit definition to list the units eligible for sampling. This seems reasonably straightforward for single-family houses, townhouses, and apartments. However, embedded household units, i.e., separate households within a single structure, such as a basement apartment, are a particularly troublesome source of undercoverage error. Field representatives are also expected to determine when tents, railroad cars, lofts, and other unconventional units qualify as housing units.

Discussions with field representatives suggested to Hainer (1987) that field representatives may have incentives for not listing some units. For example, suppose a field representative discovers a basement apartment occupied by a single male. If the field representative lists this as an extra unit, obtaining an interview may be difficult if the occupant is not often home. In most survey organizations, if the field representative ends up with a noninterview for the unit, it would count as a demerit. If the field representative simply does not list the discovered unit, undercoverage will occur and probably no one will ever know.

An alternative to area listing. As an alternative to listing area segments in rural areas and other areas without numbered addresses, the Survey of Income and Education selected individual housing units from the 1970 census address registers, and the field representatives located these units on the basis of whatever information was available (name of 1970 household head, box number, map spotting). A special coverage evaluation in these rural areas and small towns using a successor check of four structures (Marks and Nisselson 1977) estimated 7 to 13 missed housing units in missed structures per 100 enumerated housing units, significantly higher than

the estimated 4.6 missed units per 100 enumerated units in missed structures in rural areas estimated in the 1970 census. The apparent poor coverage in rural areas may have been the result of structures considered to be nonresidential or unfit for human habitation in the 1970 census being occupied when the Survey of Income and Education was conducted.

A similar methodology has been used in the 1980's for the American Housing Survey, supplemented by some area sampling to pick up the added units that were believed to be a problem in the Survey of Income and Education. The unable-to-locate rate for the 1985 survey was in the 2.1 - 2.2 percent range for the Northeast and West regions outside metropolitan areas, but was much lower outside metropolitan areas in the Midwest and South regions and within metropolitan areas in all regions (Schwanz 1988a).

#### 2.2.2. Household listing errors

Of greater concern are listing errors caused by respondent reporting error, especially for people (or subunits) within interviewed housing units. In this report, the term respondent reporting error is used to refer to all errors which occur during the interview process, whether they are caused by the field representative, the respondent, vague concepts, faulty instructions, imprecise questions, or combined effects of several of these. Respondent-related reporting errors are believed to significantly reduce coverage for most household surveys.

Among housing unit surveys, the issue of the relative importance of within-unit misses to whole-unit misses has been addressed for the National Crime Survey. The National Crime Survey has about 10 percent lower coverage of persons 12 and over, and about 6 percent lower coverage of housing units than the 1980 census. Alexander (1986) argued that whole-household undercoverage is probably less than 6 percent, since smaller housing units tend to be missed more often than large units. This means that more than 4 percent of the undercoverage occurs within units and results from household listing errors. This is one of the largest single sources of coverage error identified in this report.

Within-unit error is probably more serious for blacks and Hispanics. Hainer, et al. (1988) point out that in the CPS, black female undercoverage is close to the overall undercoverage of 7 percent, but black male undercoverage is about 20 percent, suggesting that most of this undercoverage occurs within unit as a result of household listing errors.

Hainer, et al. (1988) discuss at length the ethnographic research that has been done on household survey coverage. They suggest there are two main causes of respondent reporting error resulting in missed persons, although there have not been any good quantitative studies to show that these are major causes of coverage error.

- Some people, especially black and Hispanic males, are deliberately omitted because of potential loss of household income if their presence in the household were known to authorities.
- There is a lack of correspondence between survey definitions of household residency and how people actually live.

**Motivational causes.** Hypotheses about deliberate omissions of certain people were first inferred from 1950 census Postenumeration Survey data. As part of the Bureau of the Census's research into this, Valentine and Valentine (1971) conducted a study in a predominantly black inner city community. The study consisted of matching the household rosters for 25 units that the Valentines knew well to the reports given to Bureau of the Census field representatives. The Valentines concluded that the field representatives missed 61 percent of the males over 19 years of age, and that "... practically all the significant inaccurate information came from adult females who had some reason for neglecting to mention productive men residing in their domiciles"



(Valentine and Valentine 1971). In all 15 households where men were not reported, there was significant welfare income. Many of the unreported men were also engaged in some form of illegal economic activity, e.g., the stolen goods market. The Valentines concluded that concern about losing significant income was the reason for nonreporting.

Studies by Harwood (1970) and Hainer (1987) have come to similar conclusions. Hainer, et al. (1988) summarized: "Hainer (1987) conducted intensive interviews with his long-term informants, focusing on ways to increase the perception of confidentiality.... Hainer's informants were unanimous in their view that virtually any question that was linked to anyone's name was 'too personal' and threatening.... Informants assumed that any information given to one source is shared by all others."

Lack of correspondence between survey designer's and respondent's residency concepts. Hainer, et al. (1988) discuss the effect on coverage of differences between respondents and survey organizations in their cultural assumptions about household structure. In some cases, there are fairly simple misunderstandings, which might be corrected by better question wording, more careful definitions of concepts, and additional probing by field representatives. The Bureau of the Census has conducted some experimentation in this area (Shapiro 1986). However, more coverage error seems to be caused by fundamental differences in language or behavior rather than simple misunderstandings.

Survey organizations generally attempt to assign people a usual residence based on where they live and sleep most of the time. There are some people for whom applying this definition is problematic (Hainer, et al. 1988). Hainer (1987) and other authors have suggested that many black families can be seen as "... a group of people who share membership in a domestic group, and is an exchange network. Residence is not coterminous with 'address.' People may live in the same apartment or house, or they may instead live close by (close enough for daily interaction). Family members share clothes, store them in each of the various apartment 'addresses' shared by the family, and generally eat at the address of the family household head, usually an older black woman" (Hainer, et al. 1988).

Hainer (1987) observed a pattern in his inner city study population in which males tend to leave their families' household when 15 to 17 years old. "They reappear later as husbands or other contributors to household/families, but not in a capacity that allows their presence to be formally acknowledged" (Hainer, et al. 1988).

Hainer (1987) has also noted that family members sometimes disagree among themselves about the composition of the family. "Internal household membership is a matter of sponsorship and role performance.... Members who meet the two criteria from the standpoint of the household head do not always meet with the approval of others in the family" (Hainer, et al. 1988).

Effect of household listing errors. As discussed above, there is considerable evidence of deliberate omission of males due to fear of income loss and due to the lack of correspondence between researchers' definitions and peoples' actual residence conditions. However, there is little quantitative data on the importance of these causes. Shapiro (1979) and Pennie (1990) compared March 1980 CPS and official 1980 decennial census tabulations to determine household relationships of persons missing in the CPS but interviewed in the census. Although the Valentines' found that most of the missed males were heads of households, which would suggest that the CPS should have a higher proportion of female-headed households than the census due to deliberate omissions of males, Pennie (1990) found only some evidence for this.

As part of the 1980 census coverage evaluation program, a sample of April 1980 CPS households were matched to the corresponding 1980 census households to determine if persons found in CPS were enumerated in the census. Fay (1989) looked at the reverse match to

determine if persons enumerated in the census could be found in the same household in CPS. He found that male heads are missed much less frequently than males in the three other major relationship categories, as shown in table 13.

Table 13. Estimates of percent net CPS within-household undercoverage relative to the 1980 census for males aged 25 and over by their household status (standard errors in parentheses)

Head/spouse	0.8	(.1)
Child	12.1	(1.2)
Other relative	17.8	(1.5)
Nonrelative	22.7	(1.8)

One possible explanation for Fay's results is that male household heads who may be deliberately omitted tend to be missing in both the decennial census and housing unit surveys conducted by the Bureau of the Census. In contrast, men in other relationship categories may be missed more often in surveys than in the census because of the lack of correspondence between the survey designers' and respondents' residency concepts.

Fay (1989) estimates CPS undercoverage by age, race/ethnicity, tenure, and metropolitan-area status, as well as by household relationship status. He also fits several logistic regression equations which indicate coverage does not vary much by race/ethnicity when these other variables are taken into consideration.

There have been few studies to determine the effects of coverage problems on survey data. Thus, much of the following discussion on the effects of respondent reporting error is tentative and speculative. The discussion is taken from Hainer, et al. (1988) and Shapiro and Kostanich (1988).

One obvious potential effect is for household composition data for black and other minority groups to be biased. Shapiro and Kostanich (1988) report: "in comparison with Census Bureau interviewers, Valentine and Valentine (1971) concluded that 12% of the sampled households were female-headed vs. a Census Bureau estimate of 72%." This is inconsistent with the results of Fay (1989), which indicate that male heads of households who are reported in the census are also likely to be reported in CPS. Although the Valentines' study is small and probably an extreme case, it does suggest that surveys may be substantially overstating the number of female-headed households, especially for blacks.

Shapiro and Kostanich (1988) give adjusted estimates of black males aged 15 and over by family status. They make direct use of undercoverage rates determined by Fay (1989) from the CPS-census match discussed previously. The authors regard their estimates as speculative because Fay's results are limited to persons enumerated in the census and have other procedural limitations (see Fay 1989 for details). Biases for some categories are quite large. For example, the percentage of black male householders is estimated as 37.4 percent from the March 1985 CPS, but drops to 31.7 percent when adjusted for the effects of undercoverage.

Both deliberate omissions and misses of those with no clear usual residence probably result in significant bias for many estimates. Hainer, et al. (1988) state that deliberate omissions are "... probably extremely biasing because the reasons they are missed are so directly related to important personal and household characteristics.... For instance, Clogg, Massagli, and Eliason (1986) discuss the implausible finding from the CPS that school enrollment rates are higher for blacks than for whites, for almost every age-residence category. They speculate that this occurs because of differential undercoverage of black youth, with those attending school more likely to be counted than those who have dropped out."

As evidence of the effects of people missed because they have no clear usual residence, Hainer, et al. (1988) give this example: "... Cook (1985) presents evidence suggesting that the National Crime Survey may underestimate the number of gun assaults by as much as one-third. He offers the explanation that the National Crime Survey does not adequately cover the kinds of people criminologists believe are most likely to be involved in the life of the streets (including participation in criminal activity ...) (Cook 1985; see also Martin 1981)."

To examine the effects of coverage error on CPS data, Hirschberg, Yuskavage, and Scheuren (1977) used a different estimation method than that normally used in the survey and compared results. Their method differed from the standard method in two ways. First, it adjusted for undercoverage in the 1970 decennial census rather than controlling to census levels unadjusted for census undercount, and second, in the March supplement to the CPS, special procedures were used to assure equality between a husband's and wife's weights in addition to controlling to age-sex-race figures. The Hirschberg, et al. method was intended as an improvement over those procedures. The Hirschberg, et al. comparisons yielded substantial effects for aggregates, as would be expected, since data were controlled to larger population figures. The effects on percentages and rates were much smaller, but in some cases significant. For example, the unemployment rate increased from 4.5 percent to 5.0 percent and the poverty rate increased from 10.6 percent to 10.9 percent. Since Hirschberg, et al. were attempting complex methodology, and since there can be no assurance that their methods were correct or best, their results are difficult to interpret.

Johnston and Wetzel (1969) conducted an earlier CPS study in which they made estimates of the effect of not adjusting for the 1960 census undercount on the regular monthly estimation in the CPS. Though they found substantial effects on aggregates, they found effects on the unemployment rate to be only 0.1 percent.

Methods for reducing household listing errors. There are several means by which the number of people missed because of deliberate nonreporting or misinterpretation of concepts could be reduced. Experimentation is required to determine which of these are effective. The ideas discussed here are taken from Hainer, et al. (1988).

One approach is to change the survey household residency concept or, at least, to ask about it differently. Defining residency according to where a person is currently living or sleeping instead of where they usually live is simpler and probably less error prone for respondents. If it is undesirable or infeasible to abandon the usual residency concept, coverage might still be improved by not asking about it directly. For example, a survey might ask who slept in the unit last night, who usually eats in the unit, or even who spent any time at all in the unit yesterday. Of course, followup questions would be needed for such broad questions, as these clearly would otherwise result in overcoverage.

A second approach is to convince respondents that reporting all household residents will not adversely affect them, a quite difficult task for some population subgroups. A prerequisite for this is assuring the confidentiality of survey data and, thus, protection of the privacy of respondents. It is important to truly convince field representatives that confidentiality assurances are taken seriously, since respondents are not likely to be convinced by confidentiality assurances if field representatives themselves are at all doubtful about them. Improved training of field representatives about the subject of confidentiality is needed.

More discussion of data confidentiality and measures to protect privacy at the beginning of an interview might help allay respondents' fears. Not asking for full household rosters at the beginning of an interview may also help, especially when the questionnaire content is not threatening. Keeping a survey anonymous by not asking for names or only asking for first

names may help. Also, in face-to-face interviewing, using local, indigenous field representatives may be better, especially in low-income areas or in areas where different languages are spoken. Finally, for large-scale, frequently conducted surveys, community outreach to publicize the survey and improve public relations, as is done for the decennial census, may help.

### 2.2.3. Nonhousehold listing errors

In most nonhousehold surveys, the frame sampling unit is equivalent to the final sampling unit (see table 7). For example, in most business surveys, the establishment is both the sampling unit and the reporting unit. Occasionally, the sampling or reporting unit is the company. Coverage error is most likely to occur if a multiunit company is requested to provide data for each establishment that belongs to the population of interest. Failure to include an establishment results in undercoverage. Erroneously including an establishment results in overcoverage. However, there have been few formal studies of listing errors which result from respondent reporting errors in nonhousehold surveys.

Some coverage error resulting from respondent reporting error was found in evaluations of the 1977 Economic Censuses. Reconciliations of same-unit information were made across the Census of Retail Trade, Wholesale Trade, and Service Industries; the Current Business Surveys; and the Bureau of the Census annual business surveys. Sales (receipts) of the larger individual companies were compared among the three data sources. Millions of dollars in differences were found for many companies. About 300 of the companies with the largest differences were investigated in detail. "It was discovered that many of the firms reporting large differences were not covering the same establishments in the census and current surveys, or were not reporting as instructed. Reporting differences also resulted from different people completing the questionnaires, from dissimilar instructions, and from timing differences" (Bernhardt and Helfand 1980). It was also found that for some firms with franchised operations as well as company-owned stores, where data were requested only for the company-owned stores, duplicate sales were reported for franchised stores by the company and the franchiser. Many of the reporting differences, however, were not a result of coverage error. Reconciliation was also done between the 1977 Census of Manufactures and the Current Industrial Reports, but no coverage problems were discovered (Bernhardt and Helfand 1980).

### 2.3. Other nonsampling errors

This section includes a discussion of some other nonsampling errors which can lead to coverage problems in a survey. The first of these, recording errors, occurs when information correctly known to the survey organization is incorrectly noted. This can lead to noncoverage if the error causes incorrect exclusion or inclusion of a unit from the sample or frame. The second, responses from nonsampled units, occurs when unsampled units are substituted for sampled units in the field or when unsampled volunteers respond. In either case, the potential for biased results exists. The final part of this section presents a short discussion of unit nonresponse.

#### 2.3.1. Recording errors

A common error encountered in survey operations is recording error, that is, error arising when information correctly known is incorrectly recorded. Recording errors can be made during the execution of many different survey operations. Fortunately, the future extension of electronic data capture to many of these operations will minimize or eliminate associated recording errors. In the more typical setting, however, respondents, field representatives, screeners, data keyers, and survey analysts can all contribute to this error at various points in the survey process. Of concern here are recording errors that cause a survey unit to be dropped from or added to the sample unintentionally. Recording errors most adversely affect surveys from which small-area data are published, since the incidence of one inadvertent deletion or addition can more readily

affect these data. However, while no quantification of coverage recording errors has been found, it is the consensus among program managers that such errors are so rare as to be inconsequential. One reason for this is the fact that most survey programs make it a practice to review all survey deletions and verify that they should be dropped from the survey. If recording errors were the cause of a unit being dropped, the error would likely be discovered in this review. Recording errors are such random events that no systematic program check, except for this type of review, is likely to detect or prevent their occurrence.

There are conceptually three major points in the survey operation where recording errors are most likely to occur: At the time of the interview; at the time of screening, directory operations, and initial data entry; and during the review period when corrections to data records are made and files are updated. Field representative recording error is less a problem for the economic surveys considered in this report, since they are primarily mail-out/mail-back surveys with little or no field representative intervention. However, area-sample cases, delinquent list-sample cases, and certain failed-edit cases are handled by field representatives for the Current Business Surveys. Field representatives, of course, are vital in the execution of demographic surveys.

The probable reason that most recording errors do not result in coverage errors is that most such errors are errors of content rather than of omission or faulty inclusion. Thus, the fact that a field representative incorrectly records the information received about the occupants of a household does not lead to the household or its occupants being omitted, although it may lead to classification errors in subdomain tables. On the other hand, some surveys do employ a size criterion that a unit must satisfy in order to be included in the survey, so that a recording error based on this criterion can cause a unit to be deleted incorrectly. For example, the Pollution Abatement Survey (MA-200) conducted by the Bureau of the Census does not sample plants whose total employment is less than 20. If the plant is erroneously classified as out of scope for the survey due to a recording error, then this clearly is an error leading to undercoverage.

The type of recording error more likely to result in coverage error is the incorrect assignment of codes which directly determine whether a sampled unit does or does not belong to the survey's target population. Most surveys employ various coding schemes which denote the location, current status, classification, etc., of the survey units. For example, in the manufacturing surveys of the Bureau of the Census, coverage control codes are assigned in directory operations to indicate any changes in the operations of plants being surveyed. Many of these codes can result in a plant being deleted from the survey, so if a recording error results in the assignment of one of these codes, undercoverage results. Similarly, the SIC classification is most critical in determining the survey status of plants. If this code is incorrectly recorded, a unit may be dropped from the survey as out of scope. Such misrecordings are a possible, but very unusual, source of error in the Current Business Surveys, and in the industrial surveys of the Bureau of the Census. In most circumstances, the SIC code is incorrect because some procedure was followed incorrectly, or secondary information used to determine the code was incorrect, not because the correct code was misrecorded. Determining the right code but entering it incorrectly is still in the domain of rare events. To affect coverage, the recording error must be one that results in the unit being dropped from the sample.

Keying errors, even of codes, have little potential for causing coverage loss because most keying operations are subject to some form of verification either on a sample or 100-percent basis. Computer edit checks, as employed by the Quarterly Agricultural Surveys for example, reduce even further the possibility of these errors going undetected. Nor are corrections entered directly to a data set via an interactive terminal considered a likely source for error, since the vast majority of such corrections are made to characteristic data only. However, most of these interactive corrections are not independently verified, since for the most part they are made by analysts not prone to following a preset and independent verification regime.

### 2.3.2. Responses from nonsampled units

The probability of receiving a response for elements not in the sample should be zero, and in most surveys, responses for units not in the sample do not occur. But, administrative and clerical oversight can lead to unintentional additions. If a sample of housing units, establishments, or firms is drawn and then subsequently subsampled or refined by cutting selected elements, all records of the sample must reflect those changes. If through oversight, some of the data records contain the dropped sampled elements, normal conduct and processing of the survey will result in overcoverage. This happened in the 1984 Survey of Income and Program Participation. For budgetary reasons, a sample cut occurred at the fifth of eight interviews. Approximately 80 nonsampled units were interviewed because the dropped units had not been removed from the lists sent to field offices.

Although most surveys seek to avoid responses from volunteer units, some do not. In the Energy Information Administration's Form EIA-23 data system, a survey of oil and gas well operators, a volunteer is defined as a respondent whose assigned identification code was not on the selection list at the time the sample was drawn. Reporting requirements vary by size of operation, with the largest being required to file. Operating affiliates of a corporation are considered as individual operators and have selection and reporting requirements based on their size. Throughout the course of the survey, the forms and responses of volunteer respondents are monitored differently by the respondent tracking system than are those of the initially drawn sample. Some of these respondents clearly should not be in the survey. Others, however, should be in the survey but were not selected because the frame can never cover completely the rapidly changing population of interest.

The types of volunteer companies not incorporated in this survey include:

- Companies which filed a Form EIA-23 for the previous report year, but have not been selected in the current year and mistakenly think they have to file for the current report year, and
- Companies which complete a form forwarded by sampled companies because they sold or transferred operations to the nonsampled companies.

Some of the volunteers included are:

- A company which has a large operating affiliate and has not received a form, but is required to file, and
- A large operator which has not received a form but realizes through the *Federal Register*, its trade association, or some other means, that it should report.

The EIA's Form EIA-23 data system recognizes the possibility of volunteers at the outset and uses the respondent tracking system to help identify them. A respondent tracking system uniquely identifies each element sampled from the frame and follows the records for that respondent from the mailing of the survey form to clean entry on the final data set. At several points during the processing cycle, a list of identification codes of actual respondents is compared with those of companies on the original sample selection list, together with all previously identified and analyzed volunteers. Any new identification codes on the former list represent entities which are then added to the list of outstanding volunteers.

The weights on the randomly selected component of the sample from the EIA-23 data system are not adjusted. The appearance of new, previously unknown, entities triggers a series of actions which ultimately lead to an improvement in the frame for subsequent survey cycles.

If a sample were drawn from a perfect, current frame, then the theoretical effect of unidentified volunteer respondents would be to bias the survey's results upward. However, since sampling frames are not perfect, the effect of volunteers on coverage is not clear cut. Sometimes, when volunteers respond to surveys which have sampled from imperfect lists, this may lead to improved net coverage.

In other cases, the effect of erroneously including unsampled or unqualified respondents is to decrease slightly the efficiency of the survey and correspondingly increase respondent burden. However, if identified during survey operations, these respondents and the information they provide can be ignored in subsequent analysis. For instance, the 80 respondents mistakenly interviewed in the Survey of Income and Program Participation were identified and subsequently excluded from analysis. Some survey designs even deliberately provide for the potential inclusion of out-of-scope respondents when failure to do so would result in unacceptable undercoverage. One example of this is the Survey of Neurological Disorders. At the cost of some inefficiency, rules for screening in the housing unit portion of the survey permitted the inclusion of out-of-scope cases, which were later identified in subsequent medical evaluation and discarded from further analysis. This design permitted erroneous inclusions in order to avoid missing true cases, which would have occurred using tighter screening criteria.

### 2.3.3.- Coverage errors resulting from nonresponse

The definition of coverage error being used in this report is very broad and includes any source of error or mistake (other than sampling error) which contributes to the under- or overcoverage of the target population. Under this definition, the failure to elicit a response for a sampled target population unit or unit nonresponse can be viewed as another source of coverage error. However, in order to limit the size of this report, the following discussion of coverage errors as a result of unit nonresponse is relatively brief. For a more complete description of unit nonresponse as it contributes to incomplete data, methods to improve survey response, and procedures for adjusting for unit nonresponse, the reader is referred to the 3-volume report of the Panel on Incomplete Data in Sample Surveys (Madow, Nisselson, and Olkin 1983).

The Panel on Incomplete Data made a distinction between undercoverage and unit nonresponse as types of incompleteness using the following definitions.

"Undercoverage occurs if units that should be on the frames or lists from which a sample is selected are not on the lists, if units in the frame or sample are incorrectly classified as ineligible for the survey, or if units are omitted from the sample or skipped by the interviewer" (Madow, et al. 1983, Vol. 1, p. 16).

"Unit nonresponse occurs if a unit is selected for the sample and is eligible for the survey, but no response is obtained for the unit or the obtained response is unusable. There are four primary reasons for unit nonresponse in housing unit surveys:

- (1) No one is at the unit when the efforts are made to interview.
- (2) The interviewer cannot communicate with the persons in the unit, e.g., because of illness or a language problem.
- (3) Total refusal occurs or the interview is broken off by the respondent and the partial response prior to breakoff is classified a refusal.
- (4) The responses given by the unit are later classified as unusable" (Madow, et al. 1983, Vol. 1, p. 18).

There are a number of situations in which the distinction between undercoverage and nonresponse error, as defined by the Panel on Incomplete Data, is not clear. For example, a field representative conducts a set of interviews and mails the schedules to a central processing

facility, but they never arrive or arrive after the processing deadline. Or, a mail-out/mail-back questionnaire has been stripped of the identifying information required for using the data in estimation. What if, as occurred in the 1980 decennial census, all schedules for a particular area are destroyed by fire? In each of these cases, the sampled units could be classified as nonresponding under (4) above. On the other hand, the loss of sampled cases for a significant proportion of a subdomain of interest results in significant undercoverage for the area, and compensation with nonresponse adjustment methods may not be very satisfactory.

In housing unit surveys where the sampling frame is believed to cover the target population adequately, the incomplete recording of persons within households as discussed in section 2.2.2 above can also be considered a type of item nonresponse, where the item is the list of all persons in the household (Bailar 1984).

Differential unit nonresponse, especially where the probability of response is correlated with the variable of interest, is particularly likely to lead to biases. Consider, for example, the Consumer Expenditure Survey. If all sampled units with income greater than \$100,000 were not to respond, the effect on estimates of annual income and expenditures would be no different than if all units with income greater than \$100,000 were not represented in the sampling frame. Investigations into the response/nonresponse mechanism need to be made to determine if nonresponse adjustment methods are adequately compensating for the differential nonresponse or if, as a result of differential nonresponse, estimates of selected subdomains should not be made.

When general-purpose or national surveys are used to perform specialized analyses of demographic or geographic subpopulations, it is especially important to identify and control for the effects of differential nonresponse. For example, consider a survey designed to estimate and conduct analysis of the population in poverty. If nonresponse for this population is higher than for the general population, underrepresentation of the population in poverty will result. The data may also be less useful analytically than anticipated if the characteristics of the respondents in poverty differ from the characteristics of the nonrespondents in poverty. Thus, survey planners should carefully plan and analyze results from surveys of specialized populations, since unit coverage and unit nonresponse can be very specialized as well.

Sample attrition in longitudinal surveys can seriously affect coverage when the characteristics of the sampled units which drop out or refuse to continue to participate in the survey differ from those of the sampled units which continue to participate. Attrition is most consequential when the reasons for continued or noncontinued participation are correlated with the objectives of the survey. For example, there is evidence that sample attrition may be related to victimization status in the National Crime Survey (Biderman and Cantor 1984). To the extent victims drop out of the survey at a faster rate than nonvictims, the estimates of victimization from later interviews will be biased (U.S. Office of Management and Budget 1986).

In the Survey of Income and Program Participation, McArthur and Short (1986) found that the characteristics of the sampled units which continued to participate in the survey differed from those which did not. Out of roughly 25 characteristics examined, the following characteristics of all first-interview households were significantly different between those households which continued to participate through the fifth interview and those households which did not: Household monthly income, employment status, marital status, race, age, interview status, tenure, residence, relationship to household reference person, and region. Certainly survey estimates of two crucial variables, household income data and labor force status, will be affected by coverage loss due to attrition. Other characteristics being measured by the survey are probably affected as well.



In surveys conducted by mail, especially surveys of establishments, nonresponse may lead to overcoverage, since nonrespondents are usually assumed to be eligible sampled units from which a response should have been received. Because of this assumption, some form of imputation is performed, either explicitly or implicitly through weighting adjustments. If, in fact, nonrespondents are ineligible (e.g., establishments which are out of business), imputation results in overestimates in the same way that overcoverage does. A more detailed discussion of overcoverage as a result of imputation is given with respect to the Annual Survey of Manufactures in appendix A.1, section II, Deaths. The Annual Survey of Manufactures is generally, but not always, able to identify deaths at some point in time, but this may occur after current estimates can be corrected.

Another cause for unit nonresponse has been noted earlier in this report in the discussion of temporal errors (section 2.1.3). Coverage is influenced by the reference period(s) for which information is collected. In this ever-changing world, when the collection and reference periods are not identical, changes that occur during the periods from frame construction to sample selection or from sample selection to sample data collection can all influence coverage.

Nonresponse due to cross-sectional temporal coverage error occurs in both cross-sectional and longitudinal surveys. If a survey has established rules and procedures to maximize coverage, then field representatives and respondents need to interpret and implement those rules correctly. When respondents or field representatives are careless or unable to apply survey rules, coverage error occurs. For example, field representatives should locate the correct sampling unit. The unit might be a family that moved after the sample was selected. If survey rules specify that the family be contacted at the new address but no forwarding address or phone number is available, the field representative is unable to apply survey rules and undercoverage results.

An inverse situation may also occur. For instance, the composition of the family may have changed between the time of sample selection and the first interview of the sampled family unit. The field representative or respondent might purposely or inadvertently identify the existing family as the desired sampled unit. Undercoverage or overcoverage of the desired population could result.

Finally, field representatives in the Survey of Income and Program Participation attempt to trace movers by making inquiries of the persons now living at the sampled address and of mail carriers, rental agents, real estate companies, post offices, and through telephone directories. As a result of these procedures, Jean and McArthur (1984) found that about 80 percent of movers between the first and second interviews of the 1984 Survey of Income and Program Participation panel were traced. To improve upon this rate, the Bureau of the Census initiated a procedure whereby respondents who moved were asked to return a card with their new address. Unfortunately, even with this, there were no improvements in the response rates of movers because of respondent apathy (Kalton and Lepkowski 1985).

## CONCLUSION

The purpose of this report is to provide information to the Federal and nonfederal sectors about the existence and effects of coverage errors in surveys and to provide guidance on how to assess and improve coverage. Few studies were found by the authors, however, which actually measure coverage errors in surveys and even fewer which address the effect of coverage error on survey estimates. Of course, this report does not cover all Federal surveys and only minimal references to nonfederal surveys are provided. Even so, there appears to be an overwhelming dearth of studies devoted to measuring coverage errors and their effects. It is possible that more studies have been conducted than were found, but if so, they are certainly not readily available. This is almost as serious as not even researching coverage error, since the information in either case is not available for use in future survey work. The hope is that by increasing awareness, through this report and other media, of the potential for coverage errors in all aspects of survey design and implementation, methodologists will accept the responsibility of routinely employing and documenting more studies to identify and assess survey coverage error.

It is strongly recommended that survey researchers include methods to assess coverage in their research designs as a matter of course. In addition, they should routinely provide a discussion of populations covered and excluded, coverage error, and the measurable effects of this error on estimates when publishing survey findings. When it is not possible to estimate the effect of coverage error, even a qualitative statement about this effect should be made. It should no longer be acceptable to omit mention of coverage and coverage error in data products resulting from Federal surveys.

As to the seriousness of coverage error, the largest single source of coverage error for a housing unit survey cited in this report is an estimated 4-percent undercoverage in the National Crime Survey estimates of persons aged 12 and over due to within-housing-unit listing errors. On the economic side, the largest source of coverage error due to a single source cited is a 20-percent underestimate in the 1977 Economic Censuses statistic of receipts for nonemployer establishments due to misclassification.

These errors come from single sources. When combined across all sources, errors can become quite serious, even when each individual source is minor. Since we already know that single sources themselves can be significant, the overall effect of all sources of coverage error on survey products is of great concern. The best way to address this is to include studies to evaluate and improve coverage routinely into survey designs.

Because of the diversity of errors which lead to noncoverage, there are many methods for controlling or minimizing coverage error. The methods that apply to most surveys and which can lead to significant improvements in data quality are the use of multiple frames to improve coverage at the sampling stage and weighting adjustments to reduce the bias induced by coverage error at the estimation stage. For rare target populations, specialized sampling techniques should be considered.

It should be noted that the report is selective of the surveys discussed and the bibliography prepared. Any methodologist planning a new survey or evaluating coverage in a current survey should delve deeply into survey requirements and limitations. The information and references provided in the report should be a good starting point for many ideas, but should not be used as the only guides.

The importance of defining objectives clearly and planning accurately cannot be overemphasized. As discussed in chapter 1, survey methodologists need to make concerted efforts to thoroughly think through the objectives and issues of a proposed survey at the onset. Precise language and correct translation from the ideal survey to an implementable plan are vital;

otherwise, coverage is in jeopardy. Significant resources should be allocated in the conceptual and planning stages of a survey to minimize problems and misunderstandings that lead to coverage error. Survey designs should provide for evaluations and procedures to minimize coverage error wherever possible. If the planning is indeed thorough, survey products will be much more accurate and useful, which is the ultimate goal of any survey.

## **APPENDIX A. CASE STUDIES**

### **Introduction**

Each of the following seven case studies illustrates how the selected survey carried out by the Federal Government deals with one or more of the survey coverage problems discussed in this report. Virtually all the case studies include a discussion of methods and procedures used in the routine updating of the survey frame, a prime line of defense against coverage errors in a survey. Each highlights the application of one or more coverage assessment or improvement methods discussed in this report.

The first case study describes coverage in the Bureau of the Census' Annual Survey of Manufactures, emphasizing measures taken in the sample design to minimize coverage error, match-merging of several lists during the update of the frame, and classification errors arising during the survey.

The Long-term Care Survey, carried out on behalf of the Department of Health and Human Services, is used to illustrate coverage concerns arising from the use of an administrative data base as the sampling frame. This survey covers the noninstitutional population aged 65 and over which has functional limitations that impede normal daily activities. This second case study focuses upon the problems arising from classification errors, nonhousehold listing errors, and recording errors and provides a discussion of frame maintenance procedures.

The third case study emphasizes the methods used for developing and evaluating the National Master Facility Inventory, a comprehensive list of inpatient health facilities in the United States. The discussion covers the processes used to match-merge and unduplicate several source lists when updating and maintaining the inventory, as well as means of identifying and analyzing coverage problems.

The Bureau of Labor Statistics' Producer Price Index is the subject of the fourth case study. It emphasizes the criteria used and decisions involved in selecting a source data series for the frame, frame maintenance, classification errors, the effects of time and the temporal errors on the frame introduced in the survey, and the analysis and identification of coverage problems.

In the fifth case study, the Quarterly Agricultural Surveys carried out by the National Agricultural Statistics Service of the Department of Agriculture are discussed. These surveys, providing crop and livestock estimates at the State and national level, use a dual-frame design. Some of the coverage problems of frame construction and sample design associated with a dual-frame study are discussed.

The sixth case study presents a discussion of coverage problems encountered by the Department of Energy's Energy Information Administration in monitoring monthly deliveries of natural gas to industrial end users. This case study differs from the others in that the problem upon which it focuses originated from conceptual or relevance error rather than from the more common sources. The case study discusses the identification and analysis of coverage problems, sample design strategies used to minimize the coverage problems, and evaluation methods.

In the last case study, the coverage control and improvement procedures of the Current Population Survey are presented to exemplify those used when the sampling frame is the latest decennial census. Despite the high-coverage properties of the frame for the few months of the census-taking period, great care must be taken to minimize coverage error over the ensuing 10 years when numerous samples are selected from the census address lists.

## APPENDIX A.1. ANNUAL SURVEY OF MANUFACTURES (ASM)

### I. Introduction

At 5-year intervals (years ending in 2 or 7), an enumeration of the entire manufacturing establishment population of the United States is undertaken through the Census of Manufactures. The census collects a wide variety of information ranging from very general statistics common to all industries (employment, salaries and wages, costs of materials, inventories, etc.) to detailed statistics on industry-specific products produced and materials consumed. These data are published at various levels of industrial and geographic detail.

Except for the very small single-unit companies, the census is a mail-out/mail-back survey. Data for the small companies are imputed using payroll, shipments, and employment data provided by the administrative records of the Internal Revenue Service.

In intercensus years, a sample survey--the Annual Survey of Manufactures--is conducted. The Annual Survey of Manufactures collects virtually the same type of data on an annual basis as the census, but, in some cases, the data requested are less detailed. For example, in the annual survey, detailed census product codes (7-digit codes) are combined into product class codes (5-digit codes), and data are collected for this level. In general, however, the annual survey can be regarded as a mini-census. (See U.S. Bureau of the Census (1971) for a complete description of the Annual Survey of Manufactures.)

The Census of Manufactures serves as the primary sampling frame for the Annual Survey of Manufactures. Because of the time required to process, review, correct, and finalize the census, however, a 2-year lag exists between the census year and the first survey year of the new panel. For example, the Annual Survey of Manufactures panel selected from the 1987 census was first mailed for survey year 1989. The Annual Survey of Manufactures is also a mail-out/mail-back survey, but as does the census, it maintains a small imputation stratum which is neither sampled nor mailed. This stratum of small units is initially composed of the same administrative records imputed during the census. Each year during its birth processing, however, the Annual Survey of Manufactures adds additional cases to this stratum.

Both the Census of Manufactures and the Annual Survey of Manufactures are establishment (plant) based. A manufacturing establishment, in general, corresponds to a single physical location where manufacturing activity is performed. For the Annual Survey of Manufactures, the establishment serves as both the sampling unit and the reporting unit. However, a certainty company stratum is also defined whereby for certain companies, all the manufacturing plants owned by the companies are included in the panel as self-representing. These complete companies are included not because of sampling considerations, but rather for the use by other survey programs interested in company data over time. Sample estimates of level are formed using a fixed-base difference estimation methodology whereby sample estimates of change from the base (census) year to the current year are added to the census value.

At the time of selection, the Annual Survey of Manufactures panel is representative of the census frame from which it was chosen. This frame includes the manufacturing establishment population at a point in time. This population, however, is constantly changing as new plants are being constructed, existing plants are going out of business or are converting to different activities, plants are sold from one company to another, and companies merge and divest. In addition, the census itself is a less than perfect frame. Among other errors, it is likely to include plants not in the scope of manufacturing, be missing plants actually in manufacturing, and include industry-misclassified plants. Coverage problems associated with the census have been presumed to be minimal and primarily confined to the smallest plants which have little effect on most of the published totals. However, no formal evaluation of these errors has been undertaken.

It is necessary that the Annual Survey of Manufactures attempt to remain a panel representative of the true population over time. To this end, the survey designers maintain and apply a variety of panel maintenance rules. Some of these rules are intended to account for deficiencies that might have existed in the frame itself, while others take into account the evolving nature of the population. The panel maintenance rules are generally unbiased rules, that is, when applied correctly and comprehensively, they introduce no upward or downward bias to the survey estimates. As will be noted in some of the discussion to follow, however, it is not always practical to be comprehensive either in the application of the rules or in the identification of the complete population to which the rules are to be applied. In some cases, in fact, it is judged preferable to lower the mean square error at the expense of introducing some bias. This is particularly true at sub-U.S. levels, where a bias may be intentionally permitted at these levels, although not at the total U.S. level. The net result is that some slippage in the coverage of the ASM panel occurs, and, while the coverage loss is believed to be modest in nature, it has not been quantified.

## II. Panel maintenance operations

The major panel maintenance operations for the ASM are classified under six headings: Births, deaths, intercensus transfers, ownership changes, census misses, and miscellaneous. Limitations of these operations and their possible effect on the sample are discussed when appropriate.

**Births.** Conceptually, the treatment of new plants is straightforward. If all new operations in the course of a year could be identified, then a representative panel from these births could be selected and added to the existing panel. The problem lies in identifying the complete population of births. Births are recognized primarily through two procedures.

First, each year a complete list of newly issued employer identification numbers is obtained from the Social Security Administration. EIN's are required for tax purposes, so every employer with paid employees must have one. This list of new EIN's, however, has several limitations as a source of births. For one, an existing company can request a new EIN for a variety of reasons (reorganization, new partner). Therefore, the list contains some EIN's which do not represent new establishments. Secondly, the list finally obtained by the Bureau of the Census is incomplete. Although many firms request EIN's, they do not always inform the Social Security Administration of the nature of their operations. Thus, SSA is unable to assign an SIC code to many of these concerns, and they remain unclassified. This is a significant problem. In recent years, as many as 100,000 unclassified units have been identified, and it is estimated that as many as 6,000-10,000 may be in manufacturing. Thirdly, multiunit companies do not normally request new EIN's when they build new plants. Instead, they include them under EIN's previously assigned.

Units that are assigned a manufacturing SIC code are mailed a classification card intended to verify the SIC and to determine if the plant is a true birth. Plants indicating that they are new operations become part of the imputation stratum if they fall below the administrative record cutoff which defines that stratum. Otherwise, they are added on a sample basis to the ASM mail sample. Plants that are not births are assumed to represent existing plants requesting new EIN's. Whether or not these successor EIN's are included in the ASM is solely dependent upon whether they are in the ASM under their old (predecessor) EIN's. No presumption is made regarding establishments that do not respond to the classification card inquiry. To the extent that they are births, some downward bias is introduced, since they are not represented in the ASM. To the extent that they are successors, no bias results, since the predecessor EIN's are represented in the ASM.

The second procedure utilizes the Company Organization Survey of the large multiunit companies (employment of 50 or more), which among other things requests that the company list any additional plants not preprinted on the COS mail questionnaires and indicate whether the additional locations are newly built, acquired, or under construction. Births detected through the COS are either added in total to the mail portion of the panel or, depending upon their number, are sampled and added. This list, however, has its own deficiencies. For one, not all multiunits are included in the COS each year; a proportion of the small companies are included in each of the 3 years following the year after the census. Secondly, not all companies respond to the survey, and thirdly, many acquired plants are not identified as acquired and are treated as new. Finally, there is the possibility that some multiunit companies are not identified as such and, therefore, are not in the COS file. Typically, such undercoverage would result because the individual plants of the multiunit companies are identified on the SSEL as single units. (The SSEL is the Bureau of the Census' master list of economic entities which is updated each year by the COS and the EIN additions). Nonetheless, the COS multiunit list is believed to be close to complete.

In addition to these systematic methods for identifying birth establishments, new establishments may be identified during data collection from sampled units. Each sampled single-unit company is asked whether any additional plants operate at its location or whether the company owns any additional plants or is owned by someone else. If the establishment is determined to be part of a multiunit company, the additional plants may be birth establishments not previously identified.

It is important to recognize several factors affecting the identification and treatment of birth establishments for the ASM. The births identified from the Social Security Administration lists, which are received on a quarterly basis, typically represent the first three calendar quarters of the survey year and the last quarter of the previous year. Further, because of the lateness of the classification operation, no attempt is made to collect data from the sampled birth establishments for the current year. An imputed record is created using payroll data provided by the Internal Revenue Service.

Likewise, the identification of multiunit company birth plants by the COS occurs so late (the COS is conducted concurrently with the ASM) that no attempt is made to collect data from them until the following year. Since previous attempts to make reasonable imputations from data obtained by the COS have been unsatisfactory, imputations have not been made for these cases during the last several years. This has resulted in a slight downward bias in the ASM estimates. Beginning in 1989, these cases have been included in the mail survey in an attempt to collect current data from them.

Deaths. The treatment of deaths is straightforward from a conceptual viewpoint. As the ASM sample is a sample representative of the manufacturing establishment population, deaths in the ASM panel should be representative of deaths in the population. A number of practical problems arise, however, in the identification of deaths.

Perhaps the foremost problem associated with identifying deaths is exemplified in the saying "dead men tell no tales." In order for one to know that a plant has ceased operations, one has to be told. Since the forms for a given survey year are mailed in January of the following year, many of the out-of-business plants may have long since ceased operations. No one is now present to complete the questionnaire. From the ASM perspective, the plant is merely delinquent, so data for it will eventually be imputed, as for all nonresponding sampled units.

This is not of particular consequence for single-unit companies, since the data are imputed from Internal Revenue Service payroll data, which should reflect the fact that the plant has been inoperative during part of the year. The plant will still be considered active, however, and will

likely continue to receive the questionnaire until the Internal Revenue Service payroll data reach zero.

For multiunits, the imputation of deaths as delinquents may assume importance, since the imputation uses the prior year's data, which may be the data for an active, fully operating establishment. Normally, however, since multiunit companies do not go out of business overnight, the company will inform the ASM of individual plants that have ceased operations. It is still important to obtain data for the plants if they were in operation during part of the year, for otherwise zero records will result. On the other hand, when a company does indicate that a plant has gone out of business, there is no means of confirming it. For example, there is no Internal Revenue Service payroll for individual plants of multiunit companies that can be checked, as is the case for single-unit companies. When an entire company shuts down, however, the ASM is particularly vulnerable, for then it is possible that data for every plant of the sampled company will be imputed. The closing of the company will eventually be discovered through other operations, such as the COS canvass, but the timing may be such that the imputed data contribute to the ASM estimates for that year. It may in fact be another year or so before it is realized that the company is no longer operating. However, the infrequency of entire companies shutting down within a short span of time leads to the belief that this is a minor problem.

In summary, it is believed that the identification of death establishments is fairly comprehensive. The procedures safeguard against the erroneous deletion of plants thought to be out of business, since some positive evidence is required before the plant is actually deleted. The more likely source of error is the continuing imputation for plants that are no longer in operation.

Intercensus transfers (ICT's). Occasionally, the major activity of an establishment will change from nonmanufacturing to manufacturing. Usually no actual switch occurs, but rather the establishment is misclassified. Most of these incoming ICT's for the ASM originate in the retail and wholesale trade areas. The identification of ICT's is most likely to occur during census years, when the smaller establishments, those most likely to be misclassified, are most completely canvassed. Of course, the smallest establishments in the imputation stratum are not enumerated, and, therefore, are not subject to detection. Unless they are of significant size, the ICT's identified as manufacturing will not be added to the current ASM panel. Thus, some coverage loss occurs. Such establishments are included in the manufacturing census, however, and so will be part of the frame when the next ASM panel is chosen.

In intercensus years, the number of ICT's diminishes significantly, since only those plants belonging to the trade surveys are contacted routinely. This, plus the fact that not all ICT's are added to the ASM when found, results in some coverage loss in these years as well. Intercensus transfers can be regarded as a sample maintenance operation intended to account for deficiencies in the original sampling frame.

Ownership changes. Conceptually, ownership changes should result in no coverage loss. If a plant is in the ASM and is sold to another company, it remains in the ASM with the same sampling weight. A plant not in the ASM that is sold remains out of the ASM.

In practice, applying rules has not been so straightforward. In past years, it was often difficult to identify the successor and to determine if the successor was already included on the SSEL or needed to be added. This determination was difficult because information provided by the predecessor was often incomplete or inaccurate, especially for smaller plants, and the manual matching operation involved thousands of separated company records. Historically, therefore, the ASM managers took into account the practical limitations of attempting to link all ownership changes by establishing size cutoffs, below which no linkage was performed. With the introduction of interactive processing in recent years, the ability to obtain and process this



information has been enhanced, so that, beginning with the 1989 survey, linkage is being attempted across the entire mail panel.

Ownership changes occurring among the administrative record cases making up the imputation stratum present no problems. Even if two records appearing in the imputation stratum represent the same plant, the predecessor's record will reflect data up to the point of the change, while the successor's record will reflect data subsequent to the change.

Census misses. On very rare occasions, manufacturing plants are found to be missing, not only from the manufacturing population, but from the Bureau of the Census master establishment file (SSEL) as well. That is, the plant is completely missed and is not just misclassified. These mistakes are generally the result of human error, for example, misinterpreting company correspondence, using the wrong updating routine, etc., and are minimal in number. These misses, like ICT's, can be considered as deficiencies in the original frame.

Miscellaneous. As mentioned in the introductory remarks, decisions are made on the basis of mean square error considerations, which affect the representativeness of the sample at sub-U.S. levels. The most prominent example of this is the treatment of noncertainty sampled cases that produce such a different mixture of product classes in a given year that their basic industry classification should be changed. The ASM rule is not to change the industry code but to freeze it in its existing code. The argument advanced is that industry switches are rare events, and it is not likely that the other establishments this weighted case represents also changed to this same new industry code. Although bias is introduced in both industry estimates (but not in the total U.S. estimate), the variance component of the mean square error is substantially smaller than it would be if the change were allowed. Clearly, the sample is no longer truly representative of the industries involved. Certainty cases, which represent only themselves, are permitted to switch industries, but even they must satisfy a rigorous set of tests before the change is permitted. These tests are designed to prevent establishments from oscillating year to year from one industry to another. The freezing of the noncertainty SIC codes is maintained until the next census year (a maximum of 5 years hence) when the correct code can then be assigned.

## APPENDIX A.2. NATIONAL LONG-TERM CARE SURVEY (NLTCs)

### I. Introduction

The 1982 National Long-term Care Survey was designed to provide nationally representative data on the functionally limited population aged 65 and over. It was a detailed personal interview study of the population aged 65 and over who were not living in hospitals, nursing homes, or other institutions and who had functional limitations that impeded daily activities. The survey was sponsored by the Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services; the data were collected by the Bureau of the Census.

In 1984, the Health Care Finance Administration (HCFA) sponsored a follow-on to the 1982 survey, the 1984 National Long-term Care Survey. The 1984 National Long-term Care Survey was designed to provide information at a second time period for those persons covered in the 1982 survey, as well as a comprehensive picture of the population aged 65 and over in 1984. Data were again collected by the Bureau of the Census.

### II. Sampling frame

The target population for the 1982 National Long-term Care Survey sample consisted of noninstitutionalized persons aged 65 and over with limitations in activities of daily living (ADL) (eating, getting in or out of bed, getting in or out of chairs, toileting, dressing, bathing, walking around, or going outside) or limitations in instrumental activities of daily living (IADL) (meal preparation, laundry, light housework, grocery shopping, money management, taking medicine, or making telephone calls) lasting 3 months or longer. Because there was no available list of such persons, and virtually all of them were on the Medicare rolls, the list of Medicare enrollees was a useful frame from which to select the 1982 National Long-term Care Survey sample.

The frame population from which the 1982 National Long-term Care Survey sample was drawn included all Medicare enrollees in sampled geographic areas as of April 1, 1982. In most areas, a 10-percent sample of Medicare enrollees was selected from HCFA's December 1981 Health Insurance Skeleton Eligibility Write-off file. The sample was updated with a 10-percent sample of persons added to the file from January 1 through March 31, 1982. In areas with no December sample, a 50-percent sample was selected from all persons on the March file. The 50-percent sample was selected only in areas thought to require a sampling fraction greater than 10 percent. In total, a sample of 55,767 enrollees was selected. The sample was subsequently reduced in size to about 36,000, who were screened for limitations in activities. Of the 36,000, 6,393 were identified with such limitations and qualified for a detailed interview. Detailed interviews were completed for 6,088 persons.

The target population for the 1984 National Long-term Care Survey sample consisted of all persons aged 65 and over. First, all persons who reported functional limitations during the 1982 screening interview or who were not screened due to being institutionalized on April 1, 1982, and who survived to 1984 were interviewed regardless of their 1984 functional status. Second, from the original 25,541 persons who did not report functional impairments during the 1982 screening interview (and were not institutionalized), a random sample of 47 percent (12,100) was drawn and subjected to the same screening procedure as in 1982. In addition, a sample of elderly noninstitutionalized residents who turned 65 after the 1982 survey was screened so that a full cross-section of persons aged 65 and over in 1984 could be evaluated.

It should be noted that persons aged 65 and over who were not Medicare enrollees were missing from the sampling frame. Medicare entitlement has always been tied to a person's work history, or a spouse's work history. Individuals with no work history (excluding those receiving Supplemental Security Income) have not been entitled to receive Medicare and were thus

excluded from the frame. This frame undercoverage was estimated to be no more than 3.7 percent of the 65-and-over population of interest for 1982 and 1984. Undercoverage varied by age, race, and sex. Generally, undercoverage was greater for black persons less than 85 years old than for the corresponding nonblack persons; and, for either race, was greater for those less than 70 years than for those 70 years and over (U.S. Bureau of the Census 1986).

An additional source of undercoverage was the result of geographic interviewing constraints. Sampled persons who were found to have moved beyond 100 miles of any Bureau of the Census field representative were treated as ineligible for interview. The noninterview adjustment procedure did not adjust for these persons, although they could have been treated as noninterviewed eligible persons and included in this procedure. No estimate of the size of the undercoverage resulting from geographic interviewing constraints was made.

### III. Screening interviews and their effect on coverage in the NLTCs

For the 1982 National Long-term Care Survey, a brief series of questions was administered by telephone or personal visit interview to the 36,000 persons drawn in the sample from the Medicare enrollment files to screen out those persons without functional limitations. In the screening interviews, questions were asked to determine whether the sampled persons had an ADL or IADL limitation, and whether the limitation had existed or was expected to exist for 3 months.

Self-reporting of limitations during the screening interview had a tendency to create two types of errors:

- The sampled person reported a limitation during the screening interview which was not verified in the personal visit detailed interview--a false positive; and
- The sampled person did not report in the screening interview a limitation that did exist--a false negative.

In the National Long-term Care Survey, there was a 13-percent false positive rate among those persons for whom a detailed interview was completed. False positives were anticipated. The screening questions had been written to cast a broad net in order to minimize the false negatives, since false positives could be identified from information in the detailed interviews and eliminated, if desired, by the data analyst.

No attempt was made to measure the rate of false negatives inasmuch as this would have required administering the detailed interview to a sufficient number of people who did not report limitations on the screener to determine the rate, which could not have been accomplished within the budget fixed for this survey. The designers of the NLTCs thus assumed that the proportion of false negatives was negligible. In general, this type of assumption must be made with caution, for if the stratum defined to contain the group of elderly with no limitations, i.e., the negative stratum, was large, even a small proportion of false negatives among the negative stratum could have constituted a sizable proportion of a rare population, and thus result in coverage error (Kalton and Anderson 1986). It must be noted, however, that any attempt to measure the rate of false negatives in this instance is fraught with methodological problems. For example, elderly who reported no limitations during the screening interview, but did report them during a later followup interview, may not be false negatives but simply persons showing the effects of the aging process. Nevertheless, the potential for coverage error does exist when using a screening interview to identify a rare population, such as the target population for the NLTCs.

#### IV. Response rates

In surveys such as the National Long-term Care Survey, the final interviewed sample often differs from the initial sample for two reasons: Some members of the initial sample turn out not to be part of the population of interest, and some of the persons selected for the sample cannot be interviewed.

Of those persons initially selected for the National Long-term Care Survey, 2,472 were not in the population of interest on April 1, 1982. Some were deceased, some were institutionalized, and some lived outside of the country. A rate of ineligibility of 6.9 percent indicated that the administrative records and techniques used to provide a list from which to draw the initial sample were relatively accurate.

Of those persons who were in the population of interest on March 31, 1982, 785 had left it before they could be given the screening interview, and an additional 622 persons could not be interviewed for one reason or another. These reasons were: They could not be located; they had moved outside of any geographical area where interviewing was conducted; they were temporarily away from home or unable to respond and no proxy was available to be interviewed in their place; they refused to answer questions; or a host of other reasons.

In all, almost 96 percent of the eligible population in the sample were interviewed.

#### V. Sampling weights and their effect on coverage

Researchers in the field of long-term care research (Spillman 1989) believe problems exist with the 1984 National Long-term Care Survey cross-sectional weights on the public-use tape, resulting in a slight undercount of individuals in the community and a slight overcount of individuals in institutions. Specifically:

- The criteria used to classify individuals as institutional in the 1984 National Long-term Care Survey defined a more restricted population than the population represented by the control total used in ratio estimation. To be classified as institutional in the National Long-term Care Survey, a sampled person's residence had to have at least three unrelated residents and a health professional on duty 24 hours a day. Those who met the criteria for institutionalization, but who did not meet the more restrictive National Long-term Care Survey criteria, were classified as noninstitutional and included in the ratio estimation to noninstitutional control totals. However, the control totals were based on the more encompassing decennial census definition of institutionalization.
- Persons in correctional institutions were explicitly ineligible for the National Long-term Care Survey, but the control total was not adjusted to reflect their exclusion. Such persons account for a minuscule share of the elderly institutional population as a whole (0.3 percent), but represent larger shares of certain subpopulations, for example, 7.7 percent of black males aged 65 to 69.
- There was no attempt to poststratify the combined institutional and noninstitutional population to an estimate of the total population aged 65 and over. When the slightly too large National Long-term Care Survey institutional control total is added to the noninstitutional population control total, the resulting total of 28.03 million persons is slightly larger than the census estimate of the total resident population aged 65 and over in 1984, 27.97 million.

Manton (1988) found fault with the traditional nonresponse adjustment procedure and its subsequent effect on the 1982-1984 longitudinal weights. The nonresponse adjustment resulted

in an undercoverage of the disabled in 1982 and an overcoverage of the disabled in 1984, because the 1982 nonrespondents, a group that included 284 persons who became institutionalized after April, 1 1982, may not have responded for health reasons. Manton's solution was to create a special set of longitudinal weights. However, the author did point out that more sophisticated weights could be created to remove additional bias in the calculation of transitional rates by adjusting further the health selection effects on the sample in 1982 by calculating weights that reproduce the appropriate life table experience.

## APPENDIX A.3. NATIONAL MASTER FACILITY INVENTORY (NMFI)

### I. Introduction

The National Master Facility Inventory, established in 1962-63 by the National Center for Health Statistics (NCHS), is the most comprehensive inventory of inpatient health facilities in the United States (Science Applications International Corporation (SAIC) 1985). The NMFI program provides for the development of a list of names and addresses of all facilities or establishments within the scope of the NMFI and the collection of information which describes the facilities with respect to their size, type, and current business status (U.S. National Center for Health Statistics 1965).

The NMFI serves a twofold purpose: It is the only source of national statistics on the number, types, and geographic distribution of nursing and related-care homes, and it serves as a sampling frame for other health facility and long-term care surveys. These have included NCHS' National Nursing Home Surveys (Strahan 1987) and Early Hospital Discharge Surveys (U.S. National Center for Health Statistics 1968, SAIC 1985) and the National Center for Health Services Research's (NCHSR) Institutional Population Component of the 1987 National Medical Expenditure Survey (NMES) (Potter, Cohen, and Mueller 1987). The designs for these surveys are characterized as stratified multistage probability designs, with inpatient health facilities selected in the first stage(s) and persons selected within sampled and cooperating facilities in the last stage (Shimizu 1986; Cohen, Flyer, and Potter 1987).

The NMFI has been updated about every 2 years since 1967, using NCHS' Agency Reporting System, survey data obtained from facilities listed on the NMFI, and data obtained from State and national agencies (U.S. National Center for Health Statistics 1986). In 1985, the scope of the NMFI was expanded and data were collected under the name Inventory of Long-term Care Places. The first Inventory of Long-term Care Places survey was conducted in 1986 by NCHSR, in conjunction with NCHS and the Bureau of the Census (Potter, et al. 1987).

The scope of the NMFI inpatient health facilities has changed over time as a result of changes in the health-care industry and because new and better non-NMFI sources of data have become available. All NMFI's have included in the target population facilities defined as nursing and related-care homes. Prior to the 1978 survey, the NMFI also included hospitals (excluding Veterans Administration-operated hospitals) and other types of inpatient health facilities, such as homes for the blind, deaf, mentally retarded, and emotionally disturbed. For the 1980 survey, all identifiable hospital-based nursing homes and extended-care facilities were excluded because NCHS was unable to obtain data on all such facilities. Hospital data were subsequently obtained from the American Hospital Association. Also included for the first time in 1980 were residential community-care facilities in California and adult foster-care homes in Michigan (U.S. National Center for Health Statistics 1983). The 1982 survey added to the 1980 target population residential-care homes in Florida and Kentucky (U.S. National Center for Health Statistics 1986). For the 1986 Inventory of Long-term Care Places, the scope was further expanded to include facilities for the mentally retarded.

The first list of NMFI names and addresses was assembled in 1962 by merging a number of published and unpublished lists of hospitals and institutions (U.S. National Center for Health Statistics 1965). Sources included State license files for nursing homes and related facilities, directories maintained by national associations, and names of establishments which were contained in a subset of files maintained by the Public Health Service, the Social Security Administration, and the Bureau of the Census. NCHS then edited the list for duplicate records, and, with the assistance of the Bureau of the Census, undertook a mail survey of establishments to determine the current status, nature of business, and size of the places listed. This procedure

tended to maximize coverage and in some instances resulted in duplication on the NMFI (U.S. National Center for Health Statistics 1965).

## II. Maintaining the National Master Facility Inventory

Maintaining the NMFI involves adding the names of new facilities, deleting those that go out of business, and obtaining information on facility type and size from those currently in business. In 1964, NCHS initiated work to develop a system, known as the Agency Reporting System (ARS), by which new facilities could be identified and subsequently added to the NMFI (U.S. National Center for Health Statistics 1968). The ARS initially had 365 participating agencies: 323 State and 4 Federal agencies with authority to administer or license facilities within the scope of the NMFI, and 38 national voluntary or commercial organizations which issue lists or directories of inpatient health facilities. All 50 States and the District of Columbia were represented among the State agencies; each State averaged five ARS agencies. These agencies periodically produced lists of new inpatient facilities as part of their regular duties. This new information was then forwarded to NCHS and used to add new facilities, i.e., births, to the NMFI. However, there were a number of ARS agencies which did not identify new facilities but only provided a current list of facilities. Consequently, each new list provided to NCHS had to be matched against the most current NMFI on file to identify new facilities (U.S. National Center for Health Statistics 1968).

To update the information on each facility listed in the NMFI, a mail survey of the ARS-updated NMFI was undertaken. All surveys through 1973 were conducted by NCHS with the assistance of the Bureau of the Census (U.S. National Center for Health Statistics 1986). Survey results, data obtained through field followup to nonrespondents, and postmaster returns were used to update and edit the list of NMFI facilities by deleting those that were out of business or out of scope, adding new facilities reported by survey respondents, and modifying data on facility type and size. This periodically updated data base was used as a sampling frame and as a source of national statistics on inpatient health facilities, particularly nursing and related-care homes.

Beginning with the 1976 survey, two distinct systems were used to update the NMFI. The first system was a continuation of methods used prior to 1976. The second system, the Cooperative Health Statistics System (CHSS), decentralized NMFI data collection from the Federal to the State level. The CHSS agencies, usually State licensing agencies, were responsible for identifying new facilities and collecting updated data on existing facilities. In 1976, 16 States within CHSS collected some or all of the NMFI data. Twenty-six States participated in 1978, and by 1980, 38 States collected NMFI data. By 1982, CHSS had ceased to be active, but arrangements were made with 36 States to obtain their data for the 1982 NMFI.

The most recent NMFI update occurred as part of the 1986 Inventory of Long-term Care Places (U.S. National Center for Health Statistics 1987). The method of updating was a modification of the system used prior to 1976. Letters were sent to over 200 State and national agencies asking them to send NCHS any and all listings that they maintained for nursing and related-care homes, and facilities for the mentally retarded. Facilities not appearing on the 1982 NMFI or 1982 National Census of Residential Facilities (NCRF) (Hauber, et al. 1984) were added to form a more recent depiction of the population of interest. The 1982 NCRF was a census of residential facilities for the mentally retarded and a necessary supplement to the NMFI, since the 1982 NMFI excluded facilities for the mentally retarded, a group considered in scope for purposes of the 1986 Inventory of Long-term Care Places. A matching process was performed to remove duplicates from within and between the two files. If there were any doubts as to whether a place was a duplicate, it was retained on the Inventory of Long-term Care Places. This procedure tended to maximize coverage, but its inclusiveness resulted in duplication on the Inventory of Long-term Care Places. The Bureau of the Census then conducted a mail survey of facility administrators to obtain information on current business status, type of facility, population

served, and size (i.e., numbers of beds, residents, and annual admissions). Field followup of nonrespondents was used to reduce the nonresponse bias. Survey results, data obtained through field followup, and postmaster returns were used to update the list of facilities and facility information. This updated data base was used to create the NMES sampling frame (Potter, et al. 1987).

### III. Evaluations of the National Master Facility Inventory

The National Center for Health Statistics determined the magnitude of NMFI undercoverage using a Complement Survey (U.S. National Center for Health Statistics 1965), an application of the multiple-frame survey design technique discussed in section 1.3.3. Two frames, the NMFI and an area-sample list, were used. All institutions in the sampled area were identified, their probabilities of selection determined, and then stratified by absence/presence on the NMFI. The stratum of facilities not on the NMFI was then used to make an unbiased estimate of NMFI undercoverage.

The first Complement Survey was conducted in 1962 utilizing the design of the Health Interview Survey. Nonmatched institutions from the area sample were surveyed to collect current information on type of business and period of operation. This process yielded estimates of 5 percent NMFI gross place undercoverage and 2 percent gross bed undercoverage. However, this method of estimating types of undercoverage was far from ideal because of its large sampling error and error in field identification of institutions.

A second evaluation of the NMFI occurred in 1966 and involved reconstructing the originally assembled NMFI using the newly developed ARS (U.S. National Center for Health Statistics 1968). This evaluation pointed to a major undercoverage problem of homes for the aged that provided personal care in the States that did not license this type of facility and had no regulatory authority over them. In 1966, these States were identified as Connecticut, Idaho, Kansas, Nebraska, South Carolina, and West Virginia. Undercoverage in West Virginia was estimated at 500 facilities. Undercoverage in Idaho and South Carolina was unknown, while undercoverage in all of Connecticut, Kansas, and Nebraska was estimated at less than 25 total facilities (U.S. National Center for Health Statistics 1968).

Additional Complement Surveys were conducted in conjunction with subsequent NMFI surveys (SAIC 1985, Research Triangle Institute 1981, Shimizu 1983). Shimizu reported that on the basis of the 1982 Complement Survey, 94 percent of the eligible facilities and 98 percent of the eligible beds were included in the sampling frame for the 1985 Nursing Home Survey (Shimizu 1986).

The Cooperative Health Statistics System of updating the NMFI has also been reported as a source of NMFI undercoverage (U.S. National Center for Health Statistics 1986). Among CHSS States, there have been differences in definition, scope, and timing of the NMFI surveys because of different State licensing laws. For example, the target population for the 1980 NMFI included such facilities as personal-care homes, homes for the aged, and rest homes, but, in 1980, not all CHSS States licensed these types of facilities, resulting in some undercoverage (U.S. National Center for Health Statistics 1983).

The most recent evaluation of the NMFI began in 1983 and was completed two years later (SAIC 1985). It included a State-by-State review of health facility regulations and site visits to State agencies responsible for licensing or regulating facilities. It did not attempt to determine the degree of NMFI undercoverage as the Complement Surveys did; however, the final report concluded with a recommendation to redesign the procedures by which the NMFI was maintained. It also recommended that NMFI facility inclusion criteria be broadly defined (SAIC 1985).



#### IV. The National Master Facility Inventory as a Sampling Frame: Experience from the 1987 National Medical Expenditure Survey

As previously noted, the scope of the NMFI was recently expanded and facility data were collected through the 1986 Inventory of Long-term Care Places. The Inventory of Long-term Care Places was created to serve as the sampling frame for the Institutional Population Component (IPC) of the 1987 NMES (U.S. National Center for Health Statistics 1987). The following describes some of the coverage issues associated with the use of the NMFI and the Inventory of Long-term Care Places as sampling frames, and the methods used to correct for the potential bias. The examples rely on the NMES experience.

The NMES IPC was established to provide an assessment of the utilization, costs, sources of payment, and health status of the U.S. population in nursing and related-care homes and in facilities for the mentally retarded. The period of assessment covered calendar year 1987, during which data were collected from a sample of residents in nursing and related-care homes and in facilities for the mentally retarded. The IPC utilized a stratified three-stage probability design, with facilities selected in the first two stages and persons sampled at the last stage (Cohen, et al. 1987, Potter, et al. 1987).

Record linkage and identification of duplicates. As noted previously, the process of updating the NMFI and the Inventory of Long-term Care Places was dependent on record linkage techniques to assemble the list of facilities. These techniques may have resulted in undercoverage when facilities were erroneously classified as duplicates or overcoverage when facilities were not properly identified as duplicates. The adopted techniques minimized undercoverage at the cost of potential overcoverage (U.S. National Center for Health Statistics 1965, 1987). Five methods were used, with respect to the NMES, to correct for this source of potential bias.

- Incorporated into the instrument for the Inventory of Long-term Care Places mail survey were requests to facility administrators to return any duplicate questionnaires received under different names and/or addresses. Of the 56,720 facilities listed on the Inventory of Long-term Care Places mailing list, the Bureau of the Census classified 2,371 as duplicates on the basis of this respondent information (Potter, et al. 1987). It should be noted that the Bureau of the Census did not validate this information, and that duplicates may have been misreported by respondents and may serve as a potential source of bias.
- Results of the Inventory of Long-term Care Places mail survey were used to create the NMES sampling frame. Part of this work involved an evaluation of the Inventory of Long-term Care Places data to determine if the Inventory contained any duplicate listings. Computer-matching techniques and a visual review of the Inventory of Long-term Care Places data on facility name, address, type, and size resulted in the identification of an additional 1,570 duplicates (Potter, et al. 1987).
- After the NMES sample of 1,714 facilities was selected, a telephone screening operation was conducted to verify facility name and address prior to the first phase of field operations. A few additional duplicates were discovered when respondents indicated that they had already talked to somebody about the survey. Each case was reviewed to assure that sampled units were not erroneously classified as duplicates.
- The results of the screening revealed some remaining potential for duplication. A thorough review of the sample was then conducted to identify any potential duplicates. The potential duplicates, known as the problem pairs or triplicates, were flagged for special handling in the field. Each problem pair was assigned to the same field representative along with specific hand-written instructions on obtaining the information

necessary to resolve the problem. This information was then telephoned to a survey statistician at the home office, who determined whether there was a duplicate or two operating units at the same location.

- To assure that the above procedures resulted in the correct identification of duplicates associated with the NMES sample, an additional field procedure was developed and put into operation during the second round of NMES IPC fieldwork. An instrument known as the Duplication Worksheet was developed to identify and classify, for all NMES sampled units, all potential duplicates listed on the Inventory of Long-term Care Places list. A Duplication Worksheet listing all potential duplicate units previously classified as duplicates was created for each sampled unit. Field representatives asked facility administrators if the listed place was the same place as the sampled unit, a previous name and/or address of the sampled unit, a place affiliated with the sampled unit through administrative operating procedures or ownership, or some other place. Field representatives also inquired about any other names and addresses used by the sampled unit. The results from the Duplication Worksheet were used to adjust control totals during facility-level post-stratification adjustments in the NMES IPC.

**Listing errors.** The design of the NMES IPC called for the random sampling of current residents and admissions within sampled and cooperating facilities. Field representatives were responsible for listing and sampling persons by exacting procedures. These processes were subject to error and the following methods were used to reduce the error.

- Extensive training of the field representatives and supervisory staff was conducted.
- A programmable calculator was used to select the random sample of current residents and admissions. The programmable calculator was much easier for the field representatives to use than older methods of sample selection and, therefore, reduced the potential for error.
- Built into the programmable calculator was a review function that required field representatives to review all data entered into the calculator, thereby reducing the chance of calculator input error.
- Field representatives were instructed to call the Washington office if problems were encountered with listing or sampling. This procedure ensured that listing and sampling problems would be handled in a uniform manner at the national level rather than being subject to local variations in resolution methodology.
- Upon completion of sampling, all listing forms (except those maintained by the facility for confidentiality reasons) were forwarded to the Washington office for a 100-percent verification check. All errors in listing and sampling were reviewed by a survey statistician for corrective action.
- Ten percent of all listing, sampling, and interviewing was validated in the field by a quality control supervisor.

**Temporal errors.** The NMES IPC was designed to provide national estimates of health care utilization and expenditures for the calendar year 1987. During that year, new facilities opened, sampled facilities may have closed, sampled persons may have died, or sampled persons may have transferred out of sampled facilities. The longitudinal nature of the NMES, coupled with the length of time between the creation of the sampling frame and data collection, had implications for survey coverage. The following methods were used to control coverage.

- In the instructions for the Inventory of Long-term Care Places mail survey, facility administrators of home offices were requested to provide the Bureau of the Census with a list of names and addresses of all facilities administered by that home office. Facilities identified by this method were designated as potential births. A second list of potential births was created when Inventory of Long-term Care Places respondents returned altered copies of a single questionnaire with notations that the altered copies were for other facilities under the same administration. The two lists were compared to the original Inventory of Long-term Care Places and those facilities not appearing were added to the frame.
- Some of the NMES sample facilities were identified during the first field contact as home offices or administrative units. (Field representatives were trained on how to identify these administrative units.) A sampling procedure was developed whereby field representatives listed all units, by size and type, administered by the home office. This information was reviewed by a survey statistician and compared to the Inventory of Long-term Care Places. NMES eligible units not on the Inventory of Long-term Care Places sampling frame were identified and combined to form "super units" for the final-stage sampling process and were added to the frame.
- Closed facilities (facility deaths) were identified by postmaster returns during the Inventory of Long-term Care Places mail survey, or by field representatives calling State or local licensing agencies or local associations of health-care providers. All such NMES facilities were verified by field supervisors and classified as out of scope for the remainder of the survey.
- Sampled persons who died during the course of the survey were classified as in scope until the date of death.
- Sampled persons residing in sampled facilities that subsequently closed during 1987 were not classified as out of scope, but were followed throughout the year using the NMES IPC Survey of Next-of-Kin.
- Sampled persons who transferred out of a sampled facility to other in-scope facilities were followed to the new facility, while sampled persons who transferred back into the community were followed using the Survey of Next-of-Kin.

## APPENDIX A.4. PRODUCER PRICE INDEX (PPI)

### I. Introduction

In 1976, the Bureau of Labor Statistics commenced work on a major revision of the Wholesale Price Index (WPI). The WPI was a commodity-based index consisting of a judgmentally-selected sample of major producers. Due to the manner in which the companies and specific products were chosen, the actual coverage of the WPI fell short of its targeted population. One of the major goals of the revised index, now referred to as the Producer Price Index, was to achieve more complete coverage of the target population.

The Producer Price Index measures over-time changes in the selling prices received by goods producers from whoever makes the first purchase. Data collection procedures include a one-time personal visit to elicit cooperation and identify those products which are to be priced and subsequent monthly mailings to collect current prices for these same products. The many indexes produced each month are organized by SIC. (See U.S. Bureau of Labor Statistics (1988) for a complete description of the Producer Price Index survey design.)

For a given industry (SIC), the target population is the complete collection of transactions made by establishments classified in the SIC. An establishment is classified in the SIC in which it generates the most revenue. Sampling is done in two stages. The first stage consists of a sample of business establishments. The second stage consists of a sample of goods or services that are sold by the selected business establishments. In this case study, only the survey coverage problems which occur in the first-stage sample will be discussed.

The Producer Price Index uses a frame which is partitioned by 4-digit SIC. The UI file is the primary data source for constructing this frame. Data from secondary sources, such as trade journal publications, previous sample data, telephone calls to the businesses themselves, and interviews with industry authorities are used to augment the UI data.

### II. Evaluation and selection of primary data source

In the early development of the design of the Producer Price Index, various data sources were evaluated and compared in an effort to find the best primary source file for this survey. These potential frames were evaluated using completeness and accuracy of auxiliary information criteria (see section 1.2.1). The completeness criterion required that the frame contain all establishments currently engaged in the wholesale selling of goods and services produced at or by the establishment, regardless of the size of the establishment. The accuracy of auxiliary information criterion was based upon both the correctness of the SIC code and the adequacy of a measure of size.

The Producer Price Index publishes indexes at the industry level corresponding to the 4-digit SIC code. In order to serve as a primary source frame, establishment data must include the primary type of business. Ideally, this classification of type would be the SIC code. At a minimum, sufficient data must be present to classify each establishment by SIC code. The most desirable measure of size for the Producer Price Index is the most recent annual value of shipments and receipts for each frame unit. Short of this, reasonably accurate estimates are required. In a departure from usual establishment sampling practice, the Producer Price Index samples groups of establishments operating as a single economic unit, where the establishments are each classified in the same SIC and the prices they charge their customers are set centrally.

To satisfy the criteria given above, a single snapshot of the population frame at a given time is less useful than a source frame that is periodically updated or replaced, since establishments can

grow or shrink in size, change SIC, or go out of business, or new establishments can start up after the source frame's reference period.

Most data sources were quickly ruled out because of their lack of completeness. Due to the definition of the scope and type of sampling unit in the Producer Price Index, two source candidates were considered: The Dun and Bradstreet (D & B) file and the UI file.

Comparison of the Dun and Bradstreet file and the Unemployment Insurance file. During a pilot study of four SIC's, the D & B files were used to select samples in two SIC's, while the UI file was used to select samples in two other SIC's. Subsequent analysis showed the UI file to be superior to the D & B file as a sampling frame for the following reasons.

- The UI file was more complete than the D & B file. The coverage of the D & B file varied by SIC, since the file was composed primarily of companies for which a credit check had been made. Although the completeness of the UI file varied from State to State due to the differences in filing requirements, most companies were required to file a quarterly report to the State or States in which they operated. However, railroad workers, who were covered by the Railroad Unemployment Insurance Act, and family-owned and operated businesses with no outside employees were exempted from this requirement.
- Classifications of businesses by SIC code were found to be more accurate in the UI than in the D & B file.
- The D & B file was updated on an individual company basis, where the frequency of the updates depended on size and importance of the company and the number of credit checks D & B was asked to provide its users. A new UI file was received by BLS on a yearly basis and included updated establishment employment values.
- Because the UI file address was a mailing address used for tax purposes, it did not necessarily correspond to the physical location of the establishment. The D & B file address could also be incorrect, sometimes identifying the company headquarters or a mailing address. The State and county of the establishments were more accurate in the UI file, due to the reporting requirements of each State.
- The UI file was found to be a superior establishment file, while the D & B file was superior as a company file. The D & B file contained data on companies' organizational characteristics, such as central headquarters locations, divisional headquarters locations, and estimated employment and revenue for each company division.
- At the establishment level, the D & B file employment data were not as accurate as the UI file. These data were not critical to most of the users of the D & B file; therefore, rounding errors, rough estimations, and duplication of reported employment within a company were more likely to occur.

As a result of this comparison, the UI file was selected as the primary data source for the revision of the Producer Price Index. The D & B file was retained for several years, however, for use as a secondary source, particularly during the frame refinement process.

UI file coverage. Due to State UI laws, almost all establishments are required to file a quarterly report with the State in which the establishment is located. Although the filing requirements vary across States, 31 States have adopted the same requirements as the Federal Government: A quarterly payroll of at least \$1,500 and at least one paid employee during the preceding 20-week period. Establishments are added to the UI file in the following ways:

- A new establishment applies for a UI number,
- An unemployed worker files a claim and there is no record of his employer on the UI file,
- Field auditors find establishments not currently on the UI file, or
- The Internal Revenue Service furnishes the State with lists of new establishments that have filed for an EIN.

Conversely, establishments can be deleted from the UI file in the following ways.

- If an employer fails to file a quarterly report, an investigation may be made to determine if the company is still in business. If not, the establishment record is deleted from the UI file.
- An establishment notifies the UI office that it is out of business in order to avoid getting billed in the future.
- An establishment is in business but has zero employees for four consecutive quarters.

**Duplication in the UI file.** There are several ways in which the same establishment can be included in the UI file more than once. The State office may simply enter the data more than once, or occasionally, a company filing a supplemental report to correct an error in the original report will be entered as an original report. A company could also have been absorbed by another company, with each company filing a report during the same quarter.

### III. Producer Price Index establishment universe maintenance

Prior to the development of the Producer Price Index Establishment Universe Maintenance System, the UI file was replaced every year with a newer version. Frames for selected SIC's which were refined using the previous year's file were not matched to the newer UI file; thus, most of the refinements to the frame were lost. This caused numerous problems, most notably in the area of frame unit misclassification. Due to the lack of longitudinal tracking over the successive UI files, units could be included in more than one SIC's sampling frame if their SIC changed from one UI filing year to the next. Similarly, other establishments could miss being sampled because of this undetected movement from one SIC to another. Other refinements to the frame, such as corrected addresses and establishment names, were lost whenever a new UI file replaced the old as the primary source file.

The Producer Price Index Establishment Universe Maintenance System has been developed to address the problems described in sections 1.2.2 and 2.1.3 of this report through the longitudinal tracking of individual frame units. With this system, a new UI file is received and edited by BLS each year. Records on the new file are matched to the existing frame records and, in the event of a successful match, the existing frame record is updated with whatever incoming data are deemed to be more current or correct. The initial matching is done by computer, comparing the State, county, and UI number of each record. This matching is done across two files, the active or universe file and the death file. The death file contains a record of establishments discovered to be either out of business or classified in an industry which is outside the scope of the Producer Price Index. Incoming UI records which match up with records in the death file are maintained in the death file. Incoming UI records which match records in the active file are used to update these existing frame records. After all possible computer matches have been made, the remaining unmatched units in both the incoming UI file and the universe file are compared on a manual basis. Once all possible manual matches have been made, the universe frame records are updated with any information that is more current. Existing frame records which are not in the PPI sample and are unmatched, are moved to the death file, since these units are assumed to have gone out of business or to be represented by some unmatched incoming births. The remaining unmatched incoming UI records are considered births to the population. They are added to the frame. These birth establishments are assigned a special code to distinguish them from the other

frame members. During the yearly matching process, these birth records are put through the manual matching process, in an effort to match these units to the newer unmatched, incoming UI file members.

Effectiveness of the current Producer Price Index Universe Maintenance System. The current Producer Price Index sampling frame contains approximately 620,000 records. In the most recent capture process applied to the incoming UI file, roughly 560,000 records were matched automatically. The manual matching process resulted in approximately 1,500 additional matches. The remaining unmatched units were added to the existing universe data base. Research suggests that not all of these new frame units are actual births to the frame. Some of these incoming UI file records cannot be successfully matched to an existing frame record due to some combination of name, SIC code, ownership, or location changes.

For most frame members, the tracking process consists solely of the yearly capture of new UI file data. In these cases, the new UI file data are used to update the SIC code, employment, and the company name and address. For other frame records, data are obtained from an additional source, such as a telephone interview during frame refinement or a personal visit during the initiation process. Occasionally, these sources will yield conflicting establishment data, making it necessary to choose among them. In these cases, the information obtained from a personal interview is considered more reliable than data obtained via a telephone interview, while data recently obtained over the telephone are generally considered more accurate than the UI file data. Given the need to resample an SIC approximately every 6 years, it is important to retain the information that comes from sources other than the UI file, so long as these data are believed to be more accurate.

One way to judge the relative accuracy of incoming data is through the reference period variable. The reference period is a variable identifying the month and year in which data were obtained for that frame record. Incoming establishment information can only be used to update the present universe data base if its reference period is more recent than that of the existing file. By the time the incoming UI file is captured and used to update the frame, the data are approximately 18 months old. As a result of this time lag, any recent changes to an industry must be properly reconciled in the universe data base during the frame refinement process. At some point, the UI file may catch up with these changes. It then becomes necessary to reconcile the newer UI file data with the data already in the frame file. At present, there is no way of doing this as part of the automated capture and update process.

Refinement of a single Standard Industrial Classification sampling frame. In preparation for sampling a given SIC, primary and secondary sources are used to carry out a variety of refinement operations. The goal of this frame refinement process is to create a final sampling frame consisting of all economic units producing in the SIC. Primary sources are telephone or written contacts with companies in the SIC. An industry analyst telephones the largest companies in the SIC as a standard part of the refinement process. Information from secondary sources, such as data collection during the previous cycle, industry company lists, and trade journals, is generally verified by a telephone call to the company.

Frame refinement operations include adding and deleting units, transferring units to or from another SIC's frame, splitting frame units into separate establishment records, and combining establishments. Whenever an industry analyst proposes adding a unit to an SIC's sampling frame, a search is made for the unit in other sampling frames, as well as in the particular SIC frame, to see if it can be located as a separate unit or should be split off from an existing record. Refinement of name and address information and other attribute data not connected with the sampling process is postponed until after sample selection, and then performed only on selected units during sample refinement.

A frame unit can appear in more than one SIC frame in a given cycle. This is permitted when there is disagreement as to the most appropriate SIC classification, there has been a recent change in production, or the previous SIC classification is believed to be incorrect. If a sampled unit is out of scope for the SIC for which it was sampled, it will be dropped from that sample.

There are several types of automated report listings that provide assistance at key points of the frame refinement process. A one-time listing produced prior to the beginning of refinement for a given SIC lists the outcomes of all previous cycle sampled units. Transaction trail lists can be obtained at any stage which show in chronological order all the operations (adds, deletes, transfers, moves, changes) made on the frame since the beginning of refinement. Frame report listings show the units available for sample selection. Frame exclusion listings show records that are not available for sample selection because they were already selected in another SIC.

#### IV. Identifying the reporting unit

Although the sampling frame for a given SIC is refined prior to sample selection, it may contain units which have been improperly included, i.e., at least a portion of the unit did not belong to the specified SIC or constituted a separate unit on its own. For such a sampled unit, there are detailed collection procedures for associating the sampling unit with a reporting unit, so as to minimize the loss of coverage. When a sampled unit is fielded for collection, the field representative locates the unit and verifies county, name, address, employment, and production, noting reasons for discrepancies. For sampled units which are clusters of establishments, this information is verified for each establishment in the cluster. A sampled establishment is considered to be correctly identified if an establishment is located which has three of the four defining characteristics of the sampled establishment. The defining characteristics in the Producer Price Index are location, ownership, production, and clientele. If two or more of the defining characteristics do not match, a check is made to see if the sampled establishment is located elsewhere. If it is not found elsewhere, then the establishment represented by the original frame record is considered to be out of business.

Although an establishment in a given cluster may be correctly located, it still may not belong in the cluster because it is out of business, out of scope, misclassified, sold to a different company, or part of another records center of the same company. Whenever a portion of a cluster is determined to be sold, misclassified into one of a set of predetermined related SIC's, or part of another records center, the field representative forms a separate sampling unit for that portion known as a field-created sampling unit. When such a unit is formed, the original sampling unit is a cluster but does not constitute, as was supposed when it was sampled, a single economic unit operating in a single SIC. The new unit is retained in sample, provided it is in the sampled SIC or an SIC which is closely related to the sampled SIC. The purpose of forming field-created sampling units is to ensure that data are collected for all establishments that are legitimately part of the original sampled unit. Another method to minimize coverage loss is the treatment given data for sampled units or portions of sampled units that are misclassified into non-related SIC's. These sampling units are given another chance of selection in their proper SIC. The results of this process determine if data collection should proceed for this unit.

#### V. Collection feedback

As industry samples are selected, fielded, and collected, the collected data are used to maintain the Producer Price Index Establishment Universe Frame. Much of this collection feedback process has been automated, so that many updates to the frame occur without being manually entered. Collection feedback forms, filled out by Washington, DC staff during review of collected data, are used to facilitate the same types of operations to the frame that take place in frame and sample refinement. These forms are used primarily to collect data about establishment clusters and to make changes on the basis of the data. As part of the collection



feedback process, the name and address information for sampled units is used to update the frame. Also, units for which the collected and assigned SIC's differ are transferred into the appropriate SIC in the frame, and out-of-scope and out-of-business units are moved to the death file. Sampled units that are called out of business with respect to one SIC because of significant changes to defining characteristics are added again to the active universe file as births in the appropriate SIC. Whenever a field-created sampling unit is formed, the frame is updated accordingly.

## VI. Summary

Efforts to improve the accuracy and completeness of the Producer Price Index frame include:

- The continuous maintenance of a frame file,
- The yearly update of the Producer Price Index Establishment Universe file with data from the Unemployment Insurance file,
- The individual refinement of a single SIC's sampling frame, using telephone interviews, trade journal publications, and other secondary data sources,
- The capture and use of data obtained through a personal interview with individual sampled units, and
- Updating during the repricing process.

In addition, certain definitional and procedural changes help to eliminate coverage losses that occur whenever sampling units are incorrectly formed or some of their identifying characteristics (e.g., ownership, production, location, and organizational structure) have changed.

Two new methodologies are being established which will improve the quality of business identification and increase the detail of information on frame units. The first involves the creation of the BLS Universe Data Base (UDB) system, which is patterned after the Producer Price Index Establishment Universe Maintenance System. Currently, BLS maintains a computer system which combines the UI address files to create a national data base to be used by BLS surveys as a sampling frame. In early 1990, the UDB replaced this system. Improvements in the new data base will increase sampling applications and record linkage capabilities. Beginning with the operation of the UDB, a number of new data elements for reporting units will be stored. These elements will enable the system to trace both transfers of ownership and changes in configuration of reporting units, and to link records more comprehensively during file updates. State files will be provided quarterly instead of annually, and the production cycle will be shortened to allow frame users access to more current UI data.

The second major development is the creation of a new BLS Federal/State statistical enhancement project, the Business Establishment List (BEL). Since 1989, the BEL program has been working to transform UI data into an establishment-based micro-data file with work site identification information and physical location addresses. To accomplish this end, the BEL program has redefined many of the requirements for State collection and reporting of UI data to BLS. Employment and wage data from multiestablishment employers will be reported at the work site level rather than aggregated as one unit. Subunits of multiunit employers will be identified with unique 3-digit reporting unit numbers. Information on these subunits will include primary and secondary names, a work site description, and a physical location address. The increases in business identification information and level of detail on UI files should reduce frame refinement work.

## APPENDIX A.5. QUARTERLY AGRICULTURAL SURVEYS (QAS)

### I. Introduction

The Quarterly Agricultural Surveys, conducted by the National Agricultural Statistics Service, provide inventory and production estimates for crops and livestock at State and national levels. The Quarterly Agricultural Survey utilizes two frames: A list frame for sampling efficiency and an area frame for coverage completeness. The sampling unit for the list frame is a name. The sampling unit for the area frame is a parcel of land (segment). The reporting unit in both cases is all land operated by one or more persons under a single land-operating arrangement. Each calendar quarter a list sample of farm operators (75,000) is contacted by mail, telephone, or personal visit for inventory information on the land they operate. Sampled segments (16,000) selected from the area frame are also screened for farm operators (55,000). The multiple frame estimator utilized by NASS requires the matching of names between the two frames to identify those in the area frame who had no chance of selection from the list. These are referred to as nonoverlap operators. The nonoverlap domain estimate compensates for the incompleteness of the list, thereby completing coverage of the target agricultural population.

### II. List frame construction and maintenance

The purpose of the list frame for the Quarterly Agricultural Survey is to improve sampling efficiency. Names, addresses, phone numbers, and measures of size for farm operators permit stratification for more efficient sample selection and allow the use of less expensive survey methods for more efficient data collection. This list is not expected to be complete. Farming operations go in and out of business too quickly to expect to have a complete list. However, considerable gains in efficiency can be expected from utilizing a list frame containing a significant proportion of the larger operations (Vogel and Bosecker 1974).

Incompleteness of the list is not a coverage problem when the list is backed up by a complete area frame. List incompleteness contributes to lower sampling efficiency in a two-frame estimator, but use of the area frame removes coverage bias due to omission of population units from the list. Duplication in the list also lessens sampling efficiency, but one may appropriately compensate for coverage bias through the detection of duplicates in the sample (Gurney and Gonzalez 1972).

The list frame for the Quarterly Agricultural Survey was created using numerous input sources of farm operators. Many of the same operators appeared on several source lists. Therefore, two to three times as many names were brought together during creation of the list than eventually made up the list frame. Record linkage procedures based on work by Fellegi and Sunter (1969) and described by Coulter and Mergerson (1978) were used to standardize and remove duplication in the construction of the final composite list.

Master lists were built for several States in 1979, and all States were using the list frame system by 1982. After the initial creation phase, a continuous maintenance program has been in place to keep the frame current. New operations are added, those no longer operating are deleted, and the data associated with each active operation are updated as new information becomes available.

Six subsystems are utilized to facilitate management and utilization of the list sampling frame. The Source List Editor Subsystem standardizes input records into a common format, reduces matched records to a single record, identifies all components within the name and address of each record, and codes all names as an individual, partnership, or corporation. A Record Linkage Subsystem employs different linkage procedures for each class of names (individual, partnership, or corporate) to group potential duplicates by class into definite links, possible links, and non-links. The Group Resolution Subsystem codes a record to represent each linkage group.

matches across the linkage groups and classes, and produces the computer-identified possible matches for visual inspection. Duplication is kept to a minimum by removal of the computer-determined definite links and the identified duplicates among match groups.

The fourth list sampling frame subsystem is Data Select. This program determines the "best" data from among several input sources to attach to the list record. This might be the largest value, most current value, or the value coming from the best source, depending upon guidelines specified for each of the variables of interest. The Sample Select Subsystem then stratifies the list frame using control variables for many different surveys and selects multiple samples simultaneously.

Finally, the Mail and Maintenance Subsystem is a frame and sample management system to create mailing labels and/or listings, amend the list frame with transactions from surveys, create special comments or changes for specific surveys, provide a history and track all changes after sample selection, and combine or organize survey samples for special needs, such as elimination of multiple contacts with the same unit for different surveys in the same time period.

Utilization of all components of the list frame system provides the means to maximize list coverage for the agricultural variables of interest and minimize duplication within the list. Remaining undercoverage in the list is compensated for through the area frame sample, and remaining list duplication is adjusted for using information received for the sample.

The most serious problem that could befall the list frame in the Quarterly Agricultural Survey multiple-frame context would be for names from the area frame sample to somehow be added to the list. This would compromise the ability to estimate from the area frame the proportion of the population of farm operators who are not on the list. The necessity for independence between the list and area frames is emphasized in all National Agricultural Statistics Service training manuals and classes. A thorough discussion of the potential for list contamination and the consequences is given in Vogel and Rockwell (1977).

### III. Area frame construction and maintenance

The primary purpose of the area frame is to provide a probability mechanism for estimating the entire population of crops and livestock in the United States. Since all crop acreage and all livestock are physically located on land, complete representation is assured if the total land area is divided into sampling units. The description of how this is accomplished through land use stratification by State and county throughout the United States is given by Cotter and Nealon (1987).

In the Quarterly Agricultural Survey, the area sample supplements the list sample, accounting for list incompleteness, to provide full coverage of the agricultural population of interest. Typically, the area frame nonoverlap domain covers 10 to 20 percent of the total crop and livestock inventories.

Since complete coverage is a primary function of the area frame, there are a number of control practices in place to ensure all land is represented without duplication or omission. First, a premium is placed on the use of good, identifiable, permanent boundaries which can be marked on maps and photographs and recognized by field representatives at the site. Land use stratification boundaries and sampling unit boundaries are drawn to provide a clear demarcation even at the expense of some sampling efficiency if necessary.

Second, the areas defined by strata and clusters of sampling units are electronically digitized so the total for the frame can be computed and compared to the known land area for the county and State. The accumulated State area is allowed to vary  $\pm 0.5$  percent from the published area.

Both the original frame materials containing the boundaries and a graphic representation of the digitized boundaries are reviewed for completeness.

Finally, the sampled segments are also digitized to determine land area. Field representatives accumulate reported acres in each segment and compare the reported total against the digitized total. This control ensures complete coverage of each sampled segment.

#### IV. Rules of association

A given area of land may be represented in the Quarterly Agricultural Survey in several ways through the list frame as well as appearing in the area frame. The operation may have a name unto itself as well as having the name(s) of one or more operators associated with the land. Any of several partners may be sampled to provide the information requested for the same parcel of land.

To control this potential for duplication, there are several rules of association set forth in field representative instructions and in supervising and editing manuals. A list-dominant rule provides for the list frame to account for any land which may be reached through the list frame, that is, an area of land may belong to the area nonoverlap domain only if none of the names associated with the land is represented on the list.

Within the list frame, potential for duplication is controlled through priority rules governing which names associated with a given parcel of land will be considered the dominant sampling unit. All data for an operation will be associated with the list name assigned greatest priority. An operation name, if any, is given top priority because the name tends to be attached to land operated under that title for a longer period of time than the names of individual operators. This is particularly true in the case of managed land, where the operation must have its own name appear on the list to be considered overlap. The name of the hired manager is not used to determine the overlap status of the operation.

In the absence of an operation name or if the operation name is not on the list, the land area may be represented through the list frame by a combination of the names of the individuals who make up a partnership (second priority) or, finally, by the name of any individual actively considered an operator, alone or in partnership, if that individual participates in making the day-to-day decisions affecting the farming of land.

Partnership operations present a particularly difficult situation. It must first be determined that a true partnership in operating the land exists, i.e., more than one person jointly operates the land. Thus, since each person can report for that operation, a rule to account for the data only once is needed. To do this, the Quarterly Agricultural Survey utilizes a highest stratum rule. If more than one partner is on the list, the data will be accepted only from the partner in the list frame stratum with the highest number. If more than one partner belongs to the same highest stratum, the data will be divided equally between them. The procedure minimizes the division of data among sampling units and attaches the data with operators having the largest stratifying control data and highest sampling rate.

By far the largest portion of the list frame consists of names of individuals who operate their own farms. However, an individual may be involved in more than one operating arrangement. According to the established rules of association, individuals should report for each of the different land operations in which they are an active operator. For example, if a person has an individual operation and is a partner in another operation, that individual should provide a report for each operation. Each operation will then be considered separately according to the priority rules governing its representation on the list frame.

## V. Error avoidance

Quality control measures during construction of the frames and proper rules of association are only the first steps in ensuring proper coverage. The rules governing the representation of each population unit must be observed during data collection. The Quarterly Agricultural Survey makes use of written instructions, formal training, active supervision, questionnaire prompting, performance evaluations, and reinterview samples to aid and monitor field representative activities. Completed questionnaires are reviewed by office personnel and submitted to computer edit and analysis both within and between questionnaires.

Problem areas requiring a great deal of attention to minimize coverage errors in the Quarterly Agricultural Survey include:

- Obtaining all names actively associated with the sampling unit,
- Determination of the nonoverlap domain,
- Obtaining an accurate report of the total acres being operated,
- Reporting all data, regardless of ownership, on land operated, and
- Nonresponse.

Beller (1979) documented these areas of concern in "Error Profile for Multiple Frame Surveys."

If field representatives do not obtain all the names appropriate for the sampling unit, the rules of association described earlier cannot be applied properly. Errors could lead to either omission or duplication, depending upon the frame from which the unit was sampled and the status of the missing names on the list frame. Emphasis is given during field representative training, in the instruction manuals, and on the questionnaires to the importance of providing the operation name, if any, and the name(s) of all operators.

If the operation were found in an area frame sampled segment, all of the names will be checked against the list. When any of the names is found, the list frame identification number is attached to the corresponding name, and the operation belongs to the overlap domain. If the operation were sampled through the list frame, all other associated names are obtained and checked against the list.

Even if all names are available, it is not always an easy task to determine whether a name from the area frame is the same as one on the list. More than one individual may have the same name near the same location. In these cases, middle initials, telephone numbers, Social Security numbers, and other identifiers help determine true matches. Spellings may differ slightly or nicknames may have been used. Great care is taken to investigate possible matches. Even after the pairings have been made and the list identification number has been attached to the name from the area sample, another verification on a computer listing of matched names is required. Investigations into the operational application of these rules are reported by Bosecker and Kelly (1975), Hill and Rockwell (1977), and Nealon (1984).

An accurate report of the total acres operated is important to the area frame estimates of list incompleteness, i.e., the nonoverlap domain component. The importance of reported total land stems from the use of a proportional or weighted allocation of data associated with reporting units. The weight is determined by prorating whole farm data for the area nonoverlap respondent in proportion to the amount of land operated inside the sampled segment versus the total land operated. Complete coverage is achieved and duplication is avoided when the sum of a farm's land parcels across all possible area frame sampling units equals the total farm size, i.e., the sum of proportional weights equals one.

The potential for reporting error exists because the respondent may not include all types of land when providing the total farm acres. The portion of the farm inside the sampled area segment is outlined on an aerial photograph, facilitating a full accounting of the acreage. The remaining acreage in the farm is more dependent upon the respondent's concept of acreage to report. The questionnaire is designed to remind the respondent of all the acres operated, whether owned, rented, or managed, and all types of land, including woods, waste, and roads, so the land outside the sampled segment is reported comparably to the land inside.

Once the total land operated has been established, all requested data are to be reported regardless of who owns the crop or livestock commodity. This applies to both area and list sample respondents. Emphasis on this concept is required because a natural inclination of some respondents is to report only what they own. Since coverage for the Quarterly Agricultural Survey is dependent upon accounting for the variables of interest through the acres where they are located, much effort is expended to ensure compliance with the concept and accuracy in the reported data.

Nonresponse in the Quarterly Agricultural Survey varies by State but typically ranges from 10 to 20 percent. Two types of adjustments are made using data from only the respondents to make inferences for the total agricultural population. The first procedure adjusts sample sizes downward to the number of respondents by list stratum. Therefore, the assumption is made that within each list stratum, nonrespondents share the same agricultural characteristics as respondents.

Evidence that respondents average fewer head of livestock than nonrespondents is provided by Gleason and Bosecker (1978) and Crank (1979). Therefore a second approach is also used. Information is provided by the field representative through observation or secondary sources on the presence or absence of individual commodities for nonrespondents. Through imputation or summary adjustment, this information is used to associate respondent data with nonrespondents having similar operations.

A coverage problem posed by nonrespondents which is sometimes overlooked concerns the status or classification of the sampling unit as a viable operating entity. Simple adjustment to sample sizes for nonresponse assumes the same proportions of nonrespondents as of respondents are out of business. Imputation may assume a unit is in business. However, there are two main sources of nonresponse--refusals and inaccessibles. Refusals most often have the items of interest (which they do not want to report, so they refuse), while inaccessibles may be in business but unreachable or may not be found because they are out of business.

Nonrespondents in the Quarterly Agricultural Survey are coded as in or out of business using available information in the same way individual commodities are coded for their likelihood of existing on an operating unit. Units for which there is no evidence of current operation may therefore be more properly handled as zero-contribution sampling units.

## VI. Comparative analysis

Many of the commodity totals estimated through the Quarterly Agricultural Survey move through the agricultural marketing channels and are therefore amenable to comparison with administrative data. Some examples of this include slaughter data for hogs and cattle, milk production for dairy cows, crushings for soybeans, and sales of cotton. Even though all of the commodity may not be accounted for through one source or process, a limited number of possibilities affords the opportunity to construct a balance sheet to account for total production.

For example, survey measurements of soybean production for 1987 plus carryover soybean stocks in storage establish the total available soybeans for the 1988 marketing season. By

monitoring exports, soybean processing, seed use, imports, and remaining stocks in storage at the end of the cycle, a reasonable accounting of the uses for the total available soybeans can be made. Because of sampling errors in surveys and imperfections in administrative records, there will be residual or unexplained differences between supply and use. However, differences exist within reasonable limits and can be monitored over time. Problems in survey coverage become readily apparent with this use of administrative data.

Another useful source of data for comparison is the Census of Agriculture conducted by the Bureau of the Census at 5-year intervals. Operators provide inventory numbers for a specific date (December 31) and production statistics for the census year. The census has its own problems in achieving complete coverage and is of course also subject to nonsampling errors during data collection. However, the target population is the same as that of the Quarterly Agricultural Survey and differences between the two measurements lead to useful analysis for evaluating coverage.

The checks and balances which exist for the Quarterly Agricultural Survey estimates subject the results of this survey to a scrutiny by the data users which is rare among government surveys. Measurements for given dates are verifiable by subsequent events.

## APPENDIX A.6. MONTHLY REPORT OF INDUSTRIAL NATURAL GAS DELIVERIES

### I. Introduction

The Energy Information Administration publishes monthly State estimates of volumes of natural gas delivered to consumers, by major type of consumer. Form EIA-857, "Monthly Report of Natural Gas Purchases and Deliveries to Consumers," is completed monthly by a sample of firms that transport natural gas to consumers. These firms include inter- and intrastate pipeline companies and local distribution companies. An important aspect of the EIA-857 data system for this discussion is that the data are used to estimate monthly deliveries of natural gas to each of three consumer sectors: Residential, commercial, and industrial. In addition, estimates for total gas deliveries to consumers within each State and within each sector for all States are published, as is an estimated grand total for all gas delivered to consumers in the United States. These estimates are published in the *Natural Gas Monthly*. National estimates are also published in the *Monthly Energy Review*.

There have been two versions of Form EIA-857. The first version was approved for use in December 1984 by the Office of Management and Budget, and was in place through 1987. It asked for volumes of natural gas sold to consumers and for revenues derived from those sales. The second version of the form has been in place since January 1988. Unlike the former version, this form requires reporting on a custody rather than equity basis, asking for deliveries to consumers and revenues derived from that portion of the deliveries that is sold. Figures 1 and 2 illustrate the phenomena of deliveries to industrial consumers and sales to industrial consumers. The former illustrates the physical flow of gas from the well to the burner while the latter shows possible flows of ownership of that gas. It can be seen that the two need not be identical.

The frame for the initial EIA-857 survey included companies responding to either of two annual surveys: EIA-176, "Annual Report of Natural and Supplemental Gas Supply and Disposition," a custody-based form; or Form FERC-50, "Alternate Fuel Demand Due to Natural Gas Deficiencies," an equity-based form. A company was eligible for selection if its reports on either data system indicated deliveries or sales to consumers in either the residential, commercial, or industrial sectors. The Form FERC-50 was discontinued in 1986. Therefore, the frame now comprises only respondents to the Form EIA-176.

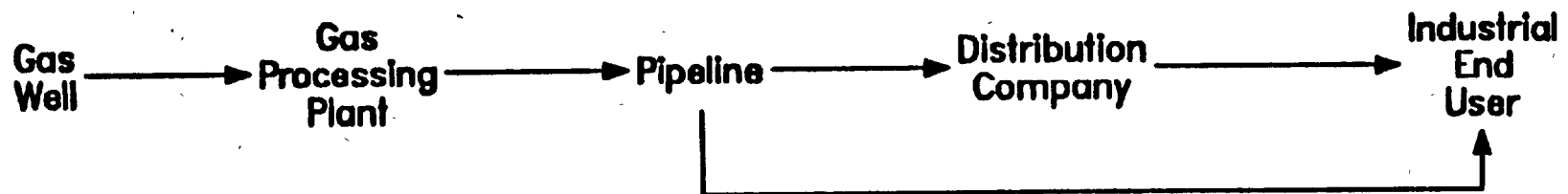
### II. Problems in coverage

At the outset, EIA recognized that there was a problem in asking for sales ("gas for which consumers were billed") rather than for total deliveries. There were some obvious logical gaps in coverage of total deliveries when measured on an equity basis. Most errors from inquiries on an equity basis arose from the fact that not all the sellers were in the frame. The frame covered all physical deliveries of natural gas. EIA knew this at the onset, and expected the effect to be trivial in the residential and commercial sectors, where the two phenomena are largely coincident. However, industrial customers were large users of fuel and had a responsibility to minimize their operating costs, so it was expected that there would be some third-party sales. Indeed, Form EIA-176 had been asking about volumes delivered to industrial customers for several years, and a noticeable volume was reported in that category.

The first solution to this problem of frame undercoverage was to adjust the monthly estimates from the survey. This adjustment was through State factors derived by comparing the EIA's official annual volumes for total deliveries to industrial consumers obtained through the EIA-176 data system to the corresponding 12-month sum of reported sales to industrial consumers. Since the monthly pattern of change in the mix of gas reported by the monthly system compared to total gas was unknown, the adjustment factor was applied consistently throughout the year.

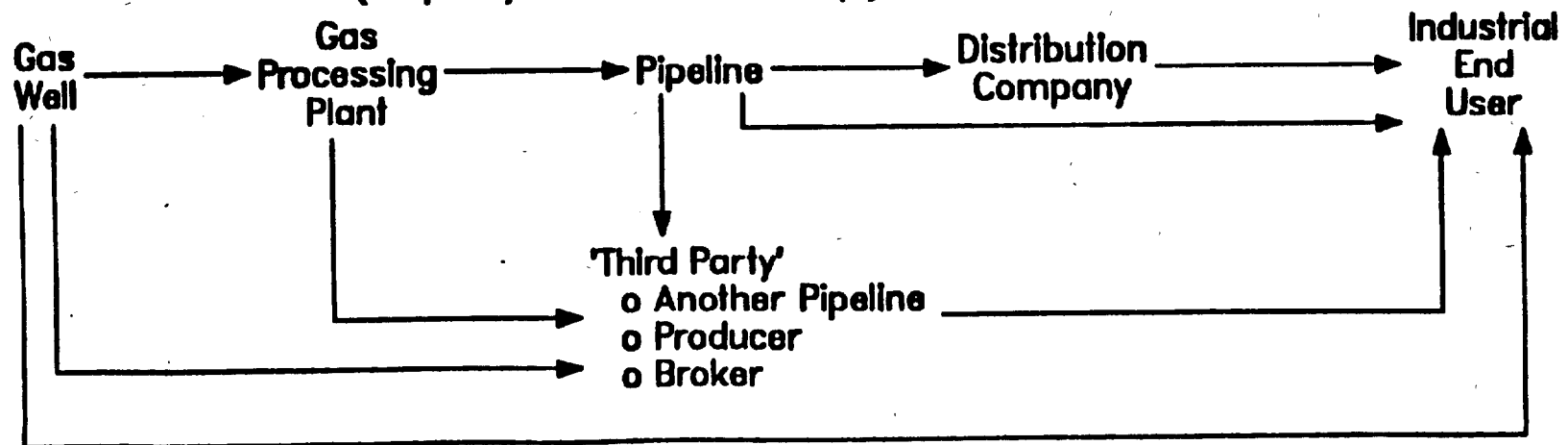


Figure 1. Typical physical flow of natural gas from gas well to industrial customer (custody relationship)



84

Figure 2. Possible financial flows (ownership) from gas well to industrial customer (equity relationship)



Published monthly volumes of deliveries to industrial consumers thus had two components: The weighted sample responses and an imputed difference based on known totals in the previous year.

After the end of a year, Form EIA-176 was used to determine the total volume delivered to consumers. Figures derived from this data system have been EIA's official volumes and prices for natural gas delivered to the residential, commercial, or industrial sectors. Accordingly, when the EIA-176 full-year totals became available, the monthly published estimates were revised. These revisions first took into account minor revisions to submissions from respondents during the year. Then the difference between the 12-month total of monthly estimates from weighted submissions and the value from the EIA-176 system was allocated among the months in proportion to the distribution of the weighted submissions.

### III. Trend in share of industrial gas transported for others

Responding to changes in the national regulatory environment and opportunities afforded by the surplus of natural gas in the mid-1980's, the natural gas industry changed the way it conducted business. More and more transactions involving large consumers were conducted through the spot market. In the spot market, a third party, the seller, was involved. No longer was there a firm relationship between the delivery and sale of gas to industrial customers. Rather, the relationship became quite fluid, with the potential existing for purchases from many sources, even within the same month. The role of the pipeline or, less often, the distribution company was becoming more akin to that of a common carrier. Their role increasingly was simply to move the commodity from one place to another.

As activity of this sort increased, it had the following impact on EIA's estimates:

- More of the sellers were out of the sample;
- More of the sellers were unknown to EIA as the importance of a new sort of function, the broker, increased; and
- The tenability of the assumptions underlying the adjustment of submitted volumes based on the previous year's experience was reduced.

By 1987, the situation had deteriorated substantially. EIA's estimates for industrial gas consumption indicated a substantial downward trend, even though gas was at a relative price advantage compared to residual fuel oil with which it competes in plants and factories having fuel switching capability. Furthermore, an increasing amount of the total gas in the national system could not be accounted for.

It was obvious that assessing the market through an equity-based data system was yielding estimates that were going increasingly awry. In reaction to these changes in the industry, EIA instituted long-term and short-term fixes. The long-term fix was to convert the monthly system from an equity to a custody basis. The short-term fix was to reevaluate the assumptions behind the monthly adjustment of weighted submissions to arrive at an estimate for publication and a change in the adjustment protocol.

Examination of the patterns in the proportion of industrial gas delivered for the account of others indicated three major facts:

- The proportion of industrial gas delivered for the account of others was increasing on a national basis;
- The pattern within States was too inconsistent to deal with adequately; and

- The assumption of a consistent relationship between reported sold volumes and true total industrial consumption throughout the year indicated by the previous year's relationship was probably damaging the estimates.

Confronted with an upward trend, EIA reevaluated the assumption behind the procedure for adjusting weighted submissions to an estimate for publication. The assumption of a constant relationship throughout the year between the two, with a sharp step implied between December and January, was discarded. Because of the inconsistent pattern within States between years, the adjustment procedure at the State level remained as it was.

The new national adjustment procedure involved two major changes. First, a linear trend in the change in the proportion of gas transported for the account of others was introduced. Second, the pivotal month for the change was shifted from January to June. In years for which there were end-of-year data (i.e., annual data for the succeeding year had been determined), the trend line was fit at both the beginning and end of the 12-month period. For 1987, when the problem was most acute, there was no end-of-year peg. The linear trend from 1985 to 1986 was allowed to continue through 1987 for purposes of the adjustment algorithm. Figure 3 shows the result of the new adjustment procedure compared to the old.

The long-term solution to the coverage problem arising from changes in the industry was a redesign of the form. As noted earlier, a new version of Form EIA-857 has been in place since January 1988. Responding to the changes in the industry, the new version has shifted from an equity to a custody basis. The change to a custody basis solved the coverage problem arising from the inability to identify companies that sold but did not deliver gas. The frame for the sample now includes respondents to the annual EIA-176 survey, which is a census of companies that deliver natural gas to end users.

EIA now believes it is more adequately monitoring the volume of natural gas being delivered to industrial consumers. Furthermore, the survey is now monitoring the relative amount that is being delivered for the transporters' own accounts and that is being delivered for the account of others.

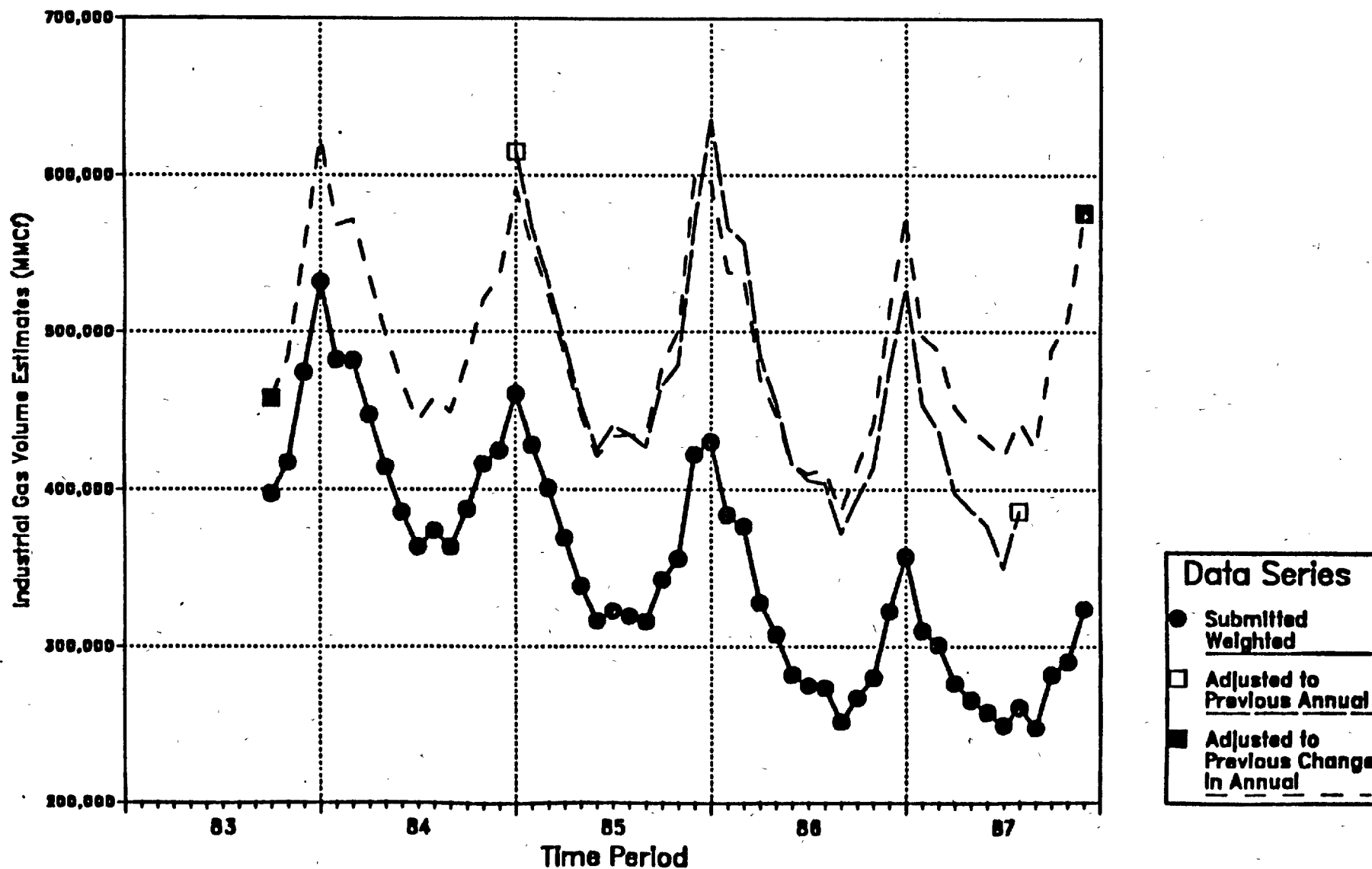
#### IV. Comparative analysis

With the introduction of the new version of Form EIA-857, reporting as it does on the same basis as required by the annual Form EIA-176, comparison of responses to identify potential coverage or respondent error problems in the monthly data system has been possible. During the summer of 1988, after 6 months of data were submitted to the monthly system and data in the annual system had been edited and cleaned, a comparison of relative volumes of industrial deliveries reported on a company and State basis was carried out. Several outliers were identified, some of which were of sufficient magnitude to substantially affect the validity of estimates derived through the data system. Followup inquiries to the companies indicated that, in a number of cases, the respondents were not reporting the volumes for which EIA had been asking. In those cases, revisions to the submissions were obtained. In other cases, the differences truly reflected changes in the volume of gas being delivered in the companies' marketing or delivery areas.

#### V. Conclusion

In this case study, some of the types of coverage error that have been encountered in the Energy Information Administration's efforts to monitor monthly deliveries of natural gas to industrial end users and the measures taken to correct them have been described. The initial version of the form, couched in terms of sales, never completely represented the population of interest because of known problems of coverage in the frame.

Figure 3. Industrial gas estimates from Form EIA-857 submissions: total United States



Analytic attempts to adjust for that undercoverage came increasingly into question as the organization of the industry changed in unanticipated ways. The eventual solution was a change to the form, and the data system, to monitor physical deliveries rather than equity transfers (sales). With this change, together with additional questions on the form asking for both sales volumes and volumes delivered for the account of others, the coverage of deliveries of gas to industrial consumers has appeared to have improved dramatically. An additional benefit of the change in the form has been that comparative analysis between the monthly and annual data system on a respondent level is now possible. This provides a means of checking for possible respondent error, as the two data systems now ask for directly comparable information, though in different time periods.

## APPENDIX A.7. CURRENT POPULATION SURVEY (CPS)

### I. Introduction

The Current Population Survey is a housing unit sample survey conducted monthly by the Bureau of the Census for the Bureau of Labor Statistics (U.S. Bureau of Labor Statistics 1989). Its primary purpose is to obtain estimates of employment, unemployment, and other characteristics of the general labor force, of the population as a whole, and of various subgroups of the population. Although Hanson (1978) and U.S. Department of Commerce (1978a) describe the sample design based on the 1970 decennial census, this work also applies to the current sample design, which is based on the 1980 decennial census.

The Current Population Survey sample design is a multistage stratified sample of the U.S. population. It is a State-based design that reflects urban and rural areas, different types of industrial and farming areas, and the major geographic divisions of each State. It is a rotating panel design wherein sampled units are interviewed for four consecutive months, dropped for eight months, and then interviewed for another four months. Each month, a new sample panel, one-eighth of the total sample, is introduced for the first time.

The target population for the Current Population Survey includes every person in the United States who is 15 years of age and over and is not institutionalized or in military service. Since 1967, the official tabulations have been restricted to data for persons 16 years of age and over. Institutionalized persons are those in correctional and health care facilities who are in the custodial care of someone else and are not free to come and go as they choose. Military persons are those who are on active duty in the Armed Forces. The target population is not restricted to citizens of the United States. Any person residing in the United States at the time of interview is a member of the population of interest. "The United States" refers to the 50 States and the District of Columbia. However, persons who are living on the premises of embassies of foreign countries, and persons who are citizens of foreign countries and are merely visiting or traveling in the United States are excluded.

### II. Sample design

To reach the target population, a sample of residences that could be occupied by persons who are not institutionalized or in the military is first selected. The primary frame for the current sample is the file of addresses created for the 1980 Census of Population and Housing.

Primary sampling unit (PSU). Before any sampling takes place, the entire United States is partitioned into basic geographic units of sampling. Traditionally, these basic units have been counties for all areas except New England, where the township is used. In some States, boroughs and independent cities are also used as basic units. These basic units are combined to create primary sampling units. Every basic unit is contained in one and only one PSU, and the complete list of PSU's geographically encompasses the entire United States. Each large metropolitan area is considered to be a PSU, although some are split for administrative reasons. Other primary sampling units are formed by grouping one or more adjacent basic geographic units. About 2,000 PSU's are formed out of the more than 3,200 basic geographic units in the United States.

The PSU's are grouped into strata, with the largest-population PSU's being placed in a stratum by themselves. One PSU is selected with probability proportional to population from each stratum. The Current Population Survey sample design consists of 713 sampled PSU's.

Within-PSU sampling. The primary frame for within-PSU sampling is the list of addresses created for the 1980 decennial census. First, a sample of census enumeration districts, a

geographic area containing 400 addresses on the average, is selected with probability proportional to the number of housing units or housing unit equivalents.

A housing unit is a group of rooms, or a single room, occupied or intended for occupancy as separate living quarters. Separate living quarters are those in which the occupants do not live and eat with any other person in the structure and that have direct access from the outside of the building or through a common hall. Housing unit equivalents are the noninstitutional group quarters population in the ED divided by 2.75, the average number of persons per housing unit. The noninstitutional group quarters population are those persons who live in such places as dormitories, rooming and boarding houses, communes, hotels, motels, and convents, and can be characterized as typically sharing some living arrangements. Each group quarters and any separate housing units associated with it are considered a special place.

An ED selected for the sample may be either an address (or list) ED or an area ED. An address ED is one that is in an area that issues permits to build new residential housing and contains less than 4 percent incomplete addresses (that is, post office boxes, rural route numbers, and so forth). An area ED is either one that is not in a permit-issuing area, or one that is in a permit-issuing area but contains 4 percent or more incomplete addresses.

**Address segments.** If the sampled ED is an address ED, the housing units in that ED are grouped into clusters of housing units (usually four to a cluster), so that the number of clusters formed equals the number of measures in the ED. The clusters are matched to determine the specific measures selected for the Current Population Survey. The 1980 census basic addresses (house number and street name) in a cluster are entered on listing sheets and are assigned for field interviewing as address segments. At the appropriate time, a Current Population Survey field representative visits the sampled cluster of one or more basic addresses in the address segment.

At take-all basic addresses (which the 1980 census listed as having one, two, three, four, and sometimes five housing units), the Current Population Survey field representative makes a new listing of all the housing units at the basic address and interviews all of them. Listings of take-all addresses are updated prior to the fifth month in the sample to account for any changes.

At non-take-all basic addresses (which the 1980 census listed as having more than four housing units), the field representative makes a new listing of all the housing units at the basic address, and interviews only those listed on lines of the listing sheet predetermined for the current sampled cluster. Non-take-all addresses are updated about once a year to account for any changes.

In the 1980 Current Population Survey design, address segments constitute about 60 percent of the sample.

**Special place segments.** If the sampled measure in an address ED is a group quarters measure, measures of size are computed for each special place in the ED. The sampled special places are identified and assigned for field interviewing as special place segments. The field representative lists the special place the month prior to its first interview. The regional office applies a predetermined random start and a sampling interval to the listing to select sampling units for interview. The special place listing is updated at least once a year to account for changes.

In the 1980 Current Population Survey design, special place segments constitute about 1.5 percent of the sample.

**Area segments.** In area ED's, blocks or chunks of land are identified which contain one or more measures. Measures of size are assigned to these blocks or chunks, and then sampled using a random start and sampling interval approach. A map is prepared outlining the sampled block or

chunk which is then assigned as an area segment. A month before the area segment is first to be interviewed, the field representative lists all housing units and special places in the area segment. The regional office applies a predetermined random start and sampling interval to select sampling units to be in sample. A different random start is used for each new measure (new Current Population Survey sample). Area segments are updated about once a year to account for any changes.

If the area segment is in a permit-issuing area, any new residential construction listed in the area segment is deleted from the sample, since it has a chance to be sampled from the new construction frame described below. New construction is represented in permit-issuing area segments in this way to control the variance of the size of the area segment between the 1980 decennial census and the time it is in sample. If the area segment is in a non-permit-issuing area, the new residential construction listed in the area segment is sampled, since that is its only chance to come into sample.

In the 1980 Current Population Survey design, area segments constitute about 28 percent of the sample. About half the area segments are permit-issuing and half are non-permit-issuing.

**Supplemental frame for new construction.** To represent in address ED's and area ED's the construction of residential housing that has occurred since the 1980 decennial census for areas that issue permits to build new residential housing, the building permits issued for residential housing units within the sampled PSU's are sampled. The source of building permit counts within the sampled PSU's is the Building Permits Survey conducted by the Bureau of the Census. These counts are obtained each month and form the basis for a sampling frame of new construction housing units.

Each month, Bureau of the Census field representatives visit a sample of building permit offices to copy addresses from permits for new residential construction issued that month. The addresses are added to the new construction frame with some geographic clustering to minimize interviewer travel costs. Those corresponding to current sampled measures are selected and assigned for field interviewing as permit segments.

As of April 1990, permit segments constitute about 13 percent of the CPS sample. The proportion of the sample that is located in permit segments increases with time since the 1980 decennial census. To maintain a constant sample size, reductions in the old construction part of the sample are made periodically.

### III. Magnitude of coverage errors

The Bureau of the Census measures overall coverage error monthly in the CPS by age, sex, race, and for Hispanics and non-Hispanics by comparing survey-based estimates to estimates based upon the most recent decennial census updated for births, deaths, immigration, emigration, and aging of the population. (The procedure by which the survey-based estimates are adjusted for noncoverage is given in this appendix, section VI.)

Tables 14 and 15 (from Hainer, et al. 1988) show average 1986 CPS coverage ratios for selected demographic categories. The ratio .932 in the upper left cell of the table means that CPS coverage for the total population is 6.8 percent lower than coverage in the census. (Note that a noninterview adjustment is applied before the coverage ratios are computed, so that noninterviews do not contribute to the ratios in these tables being less than 1.0.) The coverage of males is consistently lower than that of females, except for Hispanics aged 14 to 19. The group aged 20 to 24 has the lowest coverage, for both whites and blacks. Hainer, et al. (1988) state that "... overall undercoverage for black males is 17 percent worse than the census, and males 20-24 are 27 percent worse." The data in table 15 show that the undercoverage of Hispanics seems to



be even lower than for blacks. Twenty percent of all Hispanics are missed. Note also that these ratios do not account for undercoverage in the census.

Table 14. 1986 average coverage ratios by age, sex, and race for the CPS

	<u>Total 14+</u>	<u>14-19</u>	<u>20-24</u>	<u>25-44</u>	<u>45-64</u>	<u>65+</u>
Total	.932	.946	.887	.924	.935	.967
White						
Total	.939	.951	.902	.930	.941	.972
Male	.925	.950	.885	.914	.936	.946
Female	.950	.951	.919	.946	.946	.986
Black						
Total	.874	.904	.778	.856	.884	.946
Male	.833	.884	.733	.805	.861	.927
Female	.907	.924	.820	.910	.906	.956

Table 15. 1986 average coverage ratios for Hispanics by age and sex for the CPS

	<u>Total 14+</u>	<u>14-19</u>	<u>20-29</u>	<u>30-49</u>	<u>50+</u>
Total	.798	.845	.769	.808	.800
Male	.773	.870	.731	.762	.782
Female	.823	.820	.792	.853	.816

#### IV. Possible sources of coverage error

Census misses. In the 1980 CPS design, there is no process to account for basic addresses missed in the 1980 census in address ED's. This probably accounts for less than 1 percent of the CPS sample size. In area ED's, census misses have a chance to be sampled, since the CPS field representative makes a new listing of area segments as they come into sample.

Conversion from nonresidential to residential. Structures that were entirely nonresidential at the time of the 1980 census were not listed in the census. In address ED's, if they are converted to residential, e.g., lofts, they have no chance to be sampled. The new construction housing population includes only permits for whole-structure construction and not any conversions. The use of permits for conversions is not as clearly defined and systematic across building permit offices as it is for whole-structure construction. There is no good estimate of the extent of conversion of existing structures from nonresidential to residential in address ED's.

Time lag between permit issuance and entering the CPS sample. In the 1980 CPS design, there is a 7-month lag between the time a permit is issued for new construction and the time the structure has a chance of entering the CPS sample. This is the time it takes to list, key, cluster, sample, and prepare permit segment materials for the field. Thus, for a short period of time there are units in the new construction population that may not be represented. Linebarger (1975) estimated that approximately 12 percent of the units for which building permits were issued could be interviewed 4 months after the date of issuance. This is a cumulative figure, however, so most of these units could not have been interviewed during each of the 4 months.

Permit lag at the time of the 1980 census. Some housing units for which permits were issued prior to the 1980 census were not built until after the 1980 census, and, therefore, were not listed in the 1980 census. In address ED's and permit-issuing area ED's, these housing units had no

chance to be sampled from the census frame. However, net undercoverage was avoided by using a start date prior to the 1980 census for beginning the new construction frame. The start date was selected so that the expected number of housing units for which permits were issued prior to the start date but not built in time to be listed in the 1980 census (zero chance of selection for the 1980 CPS design) would be equal to the expected number of housing units for which permits were issued after the start date and were built in time to be listed in the 1980 census (two chances of selection for the 1980 CPS design). (See Statt, et al. 1981 for details.)

Illegal new construction. About 2 percent of the new construction in address ED's and permit-issuing area ED's is built without benefit of permit. These newly constructed units have no chance to be selected for the CPS sample. The Bureau of the Census in a 1964 study estimated illegal construction to be approximately 3.3 percent of all new construction (U.S. Bureau of the Census 1989). However, current illegal construction is believed to be lower than the 1964 estimate due to tighter zoning laws.

New construction of special places. Permits issued for the construction of new special places, whether entirely new or additional structures in existing special places, are not sampled. Thus, in address ED's, they have no chance to come into sample. This should contribute very little to the undercoverage, given the small proportion of the CPS sample that is in special places. In area segments, whether permit-issuing or non-permit-issuing, all special places are listed and sampled without regard to their date of construction.

Mobile homes. Individual mobile homes placed at addresses that were not listed in the 1980 census have no chance to be sampled in address ED's. Likewise, permits for new mobile home parks in address ED's, if issued, are not listed and sampled. In both permit-issuing and nonpermit-issuing area segments, individual mobile homes and mobile home parks are listed and sampled.

The Survey of Mobile Home Placements during April 1980 to August 1985 revealed an undercoverage of new mobile homes of 25 percent (Schwanz 1988b). Coverage improvement for mobile homes is being investigated for the post-1990 census redesign of the Bureau of the Census' demographic surveys. State health department and county tax office data are being evaluated for use as coverage improvement frames for some geographic areas, as are data from the Survey of Mobile Home Placements. If these two approaches are not feasible or are insufficient, then an area canvass approach will be used.

Year-built determination. In permit-issuing area segments, the field representative must determine the year each residential structure listed was built. Those built after the 1980 census have a chance to be sampled from the new construction housing frame. They must be deleted from permit-issuing area segment listings so they do not have two chances of selection. If the year built is determined incorrectly, a structure built before the 1980 census may be deleted by mistake, resulting in undercoverage. Likewise, field representatives may determine the year built when they should not (for example, at a mobile home or special place) and mistakenly delete such a unit from the sample. Limited investigations, however, indicate that field representatives just as often retain units in sample in error as delete them in error, due to the year-built procedure. The result is no net loss in coverage.

## V. Past attempts to remedy undercoverage

The following coverage improvement procedures developed by the American Housing Survey were used in the 1970 CPS design but are not being used in the 1980 CPS design.

Successor Check. This check was conducted in address ED's to improve coverage of conversions from nonresidential to residential, individual mobile homes placed at an address that

did not exist in the 1970 census, and existing structures moved to an address that did not exist in the 1970 census. The field representative started at a designated sampling unit, followed a specified path of travel, listed a string of eight housing units, and determined whether they existed at that address at the time of the 1980 census. The Bureau of the Census matched the strings of housing units against the 1970 census listing of addresses. Those not found in the 1970 census listing and not built after the 1970 census (and therefore not in the new construction frame) were added to the successor check frame. The improvement in coverage was marginal. Due to matching errors, some housing units added to the CPS sample were determined in the field to duplicate housing units selected for the CPS sample from the 1970 census frame. The field successor check procedure was both expensive and very difficult for the field representatives to apply consistently. All these factors led to dropping the successor check from the 1980 design (Montie and Schwanz 1977).

The Woodall frame. This frame was a commercial list of new mobile home parks. The Woodall company stopped collecting this information in 1975, so it could not be used as a source of coverage improvement for new mobile homes in the 1980 design.

The windshield survey. In address ED's, a frame of new mobile home parks was created as follows. A probability sample of about 200 tracts was selected in those PSU's expected to be most likely to create new mobile parks. The field representatives canvassed the tracts, listed any mobile home parks found (presumably by driving around and spotting them through the windshield of the car), and determined when the mobile home park was created. The Bureau of the Census matched the mobile home parks listed against the 1970 census listings for the sampled tracts. Those mobile home parks not found in the 1970 census listings were added to the windshield frame (Montie and Schwanz 1977). The limited scope of the frame (only 200 tracts) and the expense of the field canvassing and clerical matching prevented this frame from being added to the 1980 design. It may, however, be used again for the post-1990 census design.

Incomplete addresses. In the 1970 design, incomplete addresses were routinely deleted from the 1970 census listings prior to selecting the CPS sample. In the 1980 design, incomplete addresses are retained and given a chance for selection. Since an address ED must have less than 4 percent incomplete addresses, and very few address ED's have any incomplete addresses, very few incomplete addresses actually are selected for the CPS sample in the 1980 design. When they are, locator materials are prepared to help the field representative locate the incomplete addresses. The locator materials consist of a copy of the 1980 census listing that includes the incomplete address and a copy of the ED map. The 1980 census enumerator was supposed to spot the incomplete address and surrounding addresses on the ED map. Field representatives have had no difficulty locating incomplete addresses using these locator materials in the 1980 CPS design.

## VI. Evaluation and adjustment methods

Area vs. list frame coverage. In the post-1980 census design, the National Health Interview Survey (NHIS) is using an all-area design (Parsons and Casady 1986). That is, in ED's that would be address ED's in other Bureau of the Census surveys, blocks or parts of blocks are selected and assigned for field listing and interviewing as block segments. In ED's that would ordinarily be area ED's, area segments are formed in the usual way. The field representative canvasses the block segment and lists all its housing units and special places. Thus, the addresses listed in the 1980 census are not used. To offset the greater cost of an all-area listing approach, block segment listings are not periodically updated. It was felt that listing errors would be lower in block segments than in traditional area segments because, in address ED's, blocks or parts of blocks could be clearly defined, easily canvassed, and accurately listed. It was expected that updating would pick up mostly new construction, which already has a chance to be sampled in the new construction frame.

To evaluate the coverage of listing in block segments, the Bureau of the Census matched the listings of a subset of block segments to the 1980 census listing of addresses for the same block. A preliminary report (Waite 1989) shows that after reconciliation, NHIS block segments have an overall underlisting estimate of at least 3 percent. This compares to an underlisting estimate of 1 percent for traditional area segments, as measured by the 1985-88 coverage reinterview of NHIS area segments. It was not expected that block segment listings would be worse than area segment listings in NHIS. Perhaps the match operation does a better job of picking up underlisting than does coverage reinterview. In addition, NHIS coverage ratios in the 1980 design are no worse than in the 1970 design. Perhaps the NHIS block segment underlisting was compensated for by overlisting. The match study was a one-way match and did not measure overlisting. A new two-way field match and update study has been proposed for NHIS in 1990 to improve measurement of coverage error in the NHIS all-area design.

Adjustment methods. As stated in section III, the Bureau of the Census produces population estimates monthly by updating the last decennial census figures for births, deaths, immigration, emigration, and aging of the population. In addition to using these population estimates to measure undercoverage, as discussed in section III, they are used in a weighting adjustment in the CPS and most other housing unit surveys conducted by the Bureau of the Census. In this adjustment, the weight for each person in the sample is modified so that the CPS estimates for the population by age, sex, and race categories and for Hispanic and non-Hispanic groups agree with the independently determined population estimates. To the extent that labor force characteristics of missed persons are the same as the labor force characteristics of covered persons in the same age-sex-race group, this weighting adjustment reduces the bias caused by undercoverage. Hanson (1978) points out: "... this adjustment should be regarded as possibly ameliorating, but certainly not as removing the potential bias involved in coverage losses." Furthermore, the adjustment does not account for undercoverage in the decennial census. (See Hanson (1978) and U.S. Bureau of Labor Statistics (1989) for more details.)

In preparing the updated population estimates, the Bureau of the Census uses data on births and deaths from the National Center for Health Statistics and on military population deaths from the Department of Defense and the Coast Guard. The Immigration and Naturalization Service and three other agencies provide data on immigration. Estimates of the Armed Forces and the institutionalized population are subtracted out in order to produce estimates of the civilian noninstitutional population. All these computations are done by single year of age, by race, and by sex. (For more details, see U.S. Bureau of Labor Statistics (1989).)

## APPENDIX B. GLOSSARY OF ACRONYMS

ADL	Activities of daily living
ARS	Agency Reporting System
ASM	Annual Survey of Manufactures
BEL	Business Establishment List
BLS	Bureau of Labor Statistics
CE	Consumer Expenditure (Survey)
CES	Current Employment Statistics (Survey)
CHSS	Cooperative Health Statistics System
COS	Company Organization Survey
CPI	Consumer Price Index
CPP	Current Point of Purchase Survey
CPS	Current Population Survey
D & B	Dun and Bradstreet
DOD	Department of Defense
ECI	Employment Cost Index
ED	Enumeration district
EIN	Employer identification number
EIA	Energy Information Administration
FEA	Federal Energy Administration
HCFA	Health Care Finance Administration
IADL	Instrumental activities of daily living
ICT	Intercensus transfer
IPC	Institutional Population Component
IPP	International Price Program
IRS	Internal Revenue Service
LFS	Labor Force Survey (of Canada)
NASS	National Agricultural Statistics Service
NCHS	National Center for Health Statistics
NCHSR	National Center for Health Services Research
NCRF	National Census of Residential Facilities
NCS	National Crime Survey
NHIS	National Health Interview Survey
NLTCS	National Long-term Care Survey
NMCUES	National Medical Cost Utilization and Expenditure Survey
NMES	National Medical Expenditure Survey
NMFI	National Master Facility Inventory
OES	Occupational Employment Statistics (Survey)
OSH	Occupational Safety and Health
PPI	Producer Price Index
PSU	Primary sampling unit
QAS	Quarterly Agricultural Surveys
RDD	Random-digit dialing
R & D	Research and Development Survey
SAIC	Science Applications International Corporation
SIC	Standard Industrial Classification
SIPP	Survey of Income and Program Participation
SSA	Social Security Administration
SSEL	Standard Statistical Establishment List
TAR	Tape Address Register
UDB	Universe Data Base
UI	Unemployment Insurance
WPI	Wholesale Price Index

## APPENDIX C. GLOSSARY OF TERMS

### AREA FRAMES

A sampling frame based on lists of geographical units (Groves 1989, p. 100).

### AREA SAMPLING

"The entire area in which the population is located is subdivided into smaller areas, and each elementary unit ... is associated with one and only one such area ..." (Hansen, Hurwitz, and Madow 1953, Vol. I, p. 244).

"A method of sampling used when no complete frame of reference is available. The total area under investigation is divided into small sub-areas which are sampled at random or by some restricted random process. Each of the chosen sub-areas is then fully inspected and enumerated, and may form a frame for further sampling if desired. The term may also be used (but is not to be recommended) as meaning the sampling of a domain to determine area, e.g., under a crop" (Kendall and Buckland 1971).

### AREA SEGMENT LISTING ERRORS

Listing errors associated with area sampling (this report, p. 39).

See LISTING ERRORS and AREA SAMPLING.

### BIRTHS

Units that came into existence after frame construction.

### BUSINESS

See ESTABLISHMENT.

### CENSUS

"The complete enumeration of a population or groups at a point in time with respect to well-defined characteristics: for example, Population, Production, Traffic on particular roads. In some connection the term is associated with the data collected rather than the extent of the collection so that the term *Sample Census* has a distinct meaning" (Kendall and Buckland 1971).

"The modern population census may be defined as the process of collecting, compiling and publishing demographic, social and economic data about the population of a defined territory at a specified time ... either on a *de facto* or *de jure* basis" (Pollard, Usef, and Pollard 1974, p. 3).

### CLASSIFICATION ERRORS

"Error caused by conceptual problems and misinterpretations in the application of classification systems to survey data" (Hansen, Hurwitz and Madow 1953, Vol. I, p. 84).

An error that occurs because units as members of the target population are misrepresented as out-of-scope units. To the extent frame or sampled units are misclassified as out of the scope of the survey, undercoverage occurs. Classification error may also occur by the misrepresentation of units as members of the target population when in truth they are not. This results in

overcoverage. Classification errors are a type of rule-of-association error leading to noncoverage (this report, p. 33).

## COMPANY

A company or enterprise consists of one or more establishments under common ownership or control (U.S. Executive Office of the President 1987, p. 12).

## COMPLETE COVERAGE

"A survey (or census) should be called complete if virtually all of the units in the population under study are covered" (Moser and Kalton 1971, p. 54).

## CONCEPTUAL ERROR

"In planning a survey the purposes of the survey are made explicit. The purposes are then translated into a set of definitions of the characteristics for which data are to be collected and into a set of specifications for collecting, processing, and publishing. The possibilities of error arise where the statistician fails to understand the purposes of the survey, where the definitions that are set up may not be pertinent to the purposes, where the specifications (for the sample, the questionnaire, the method of collecting the data, the methods of selection and training of personnel, processing methods, etc.) would lead to error even if followed exactly" (Hansen, Hurwitz, and Madow 1953, Vol. I, p. 83-84).

## CONSUMER UNITS

A consumer unit comprises either: (1) all members of a particular household who are related by blood, marriage, adoption, or other legal arrangements; (2) a person living alone or sharing a household with others or living as a roomer in a private home or lodging house or in permanent living quarters in a hotel or motel, but who is financially independent; or (3) two or more persons living together who pool their income to make joint expenditure decisions. Financial independence is determined by the three major expense categories: Housing, food, and other living expenses. To be considered financially independent, at least two or three major expense categories have to be provided by the respondent (U.S. Bureau of Labor Statistics 1986, p. 46).

## CONTENT ERROR

"Errors of observation or objective measurement, of recording, ... which result in associating a wrong value of the characteristic with a specified unit. (Coverage errors are excluded from this definition.)" (U.S. Bureau of the Census (no date), p. 48).

## COVERAGE ERROR

Errors in coverage occur when target population units are missed during frame construction or sample data collection (undercoverage) or when they are duplicated or enumerated in error (overcoverage). Errors in definitions or applying a definition, as well as errors in locating a sampled unit, may affect coverage. Coverage errors may also occur if subclasses of the population have no probability of being included or inappropriate probabilities of being included in a sample. Response errors, which may occur when a respondent misunderstands a question, responds incorrectly because of a belief that an incorrect answer may increase his prestige, etc. or when an interviewer misasks a question or misrecords a response, may also result in errors in coverage (Hansen, Hurwitz, and Madow 1953, Vol. I, p. 84).

**"Undercoverage:** units (e.g., households, persons, establishments, farms) that should be in the frames (or lists) from which a sample is selected are not in those frames, or units in the sample are mistakenly classified as ineligible or are omitted from the sample or from the units interviewed" (Madow, et al. 1983, p. 3).

"The error in an estimate that results from (1) failure to include in the frame all units belonging to the defined population; failure to include specified units in the conduct of the survey (undercoverage), and (2) inclusion of some units erroneously either because of a defective frame or because of inclusion of unspecified units or inclusion of specified units more than once, in the actual survey (overcoverage)" (U.S. Bureau of the Census (no date), p. 48).

"The failure to give any chance of sample selection to some persons in the population" (Groves 1989, p. vi).

"Exists because some persons are not part of the list or frame (or equivalent materials) used to identify members of the population. Because of this they can never be measured, whether a complete census of the frame is attempted or a sample is studied" (Groves 1989, p. 11).

"Refers to the discrepancy between statistics calculated on the frame population and the same statistics calculated on the target population. Coverage error arises from failure to give some units in the target population any chance of being included in the survey, from including ineligible units in the survey, or from having some target population units appear several times in the frame population.... Coverage error is a function of both the proportion of the target population that is not covered by the frame and the difference on the survey statistic between those covered and those not covered.... Coverage error is a property of a statistic, not a survey" (Groves 1989, pp. 83-85).

See also NONCOVERAGE, OVERCOVERAGE and UNDERCOVERAGE.

#### CROSS-SECTIONAL COVERAGE ERROR

An error in a sample estimate that results from unaccounted for changes in the sample population from the time of frame establishment to the first interview. A type of temporal error that is a source of coverage error (this report, p. 36).

#### CROSS-SECTIONAL TEMPORAL COVERAGE ERROR

See CROSS-SECTIONAL COVERAGE ERROR.

#### CROSS-SECTIONAL SURVEY

A survey in which data are gathered on "a cross-section of the population at a single point in time" (Bailey 1982, p. 34).

#### DEATHS

Inactive frame elements (this report, p. 15).

A sampling unit which has been identified as out of business or out of scope (this report, p. 17).

#### DUPLICATION

See OVERCOVERAGE.



## ELEMENTS

"The elements of a population are the units for which information is sought; they are the individuals, the elementary units comprising the population about which inferences are to be drawn. They are the units of analysis, and their nature is determined by the survey objectives" (Kish 1965, pp. 6-7).

"The smallest units into which the population can be divided" (Sukhatme and Sukhatme 1970, p. 222).

"Each entity from the population that is the ultimate sampling objective is called a sampling element" (Bailey 1982, p. 85).

"An object on which a measurement is taken" (Scheaffer, Mendenhall, and Ott 1986, p. 20).

## ESTABLISHMENT

An establishment is "an economic unit, generally at a single physical location, where business is conducted or where services or industrial operations are performed; for example: a factory, mill store, hotel, movie theater, mine, farm, ranch, bank, railroad depot, airline terminal, sales office, warehouse, or central administrative office" (U.S. Executive Office of the President 1987, p. 12).

## FRAME

The frame is any material, device, etc., which is used to provide observational access to the population (Dalenius 1974).

A list of the sampling units which make up the population (Cochran 1963, p. 7).

"Physical lists and procedures that can account for all the sampling units without the physical effort of actually listing them" (Kish 1965, p. 53).

The frame consists of previously available descriptions of the objects or material related to the physical field in the form of maps, lists, directories, etc., from which sampling units may be constructed and a set of sampling units selected; and also information on communications, transport, etc., which may be of value in improving the design for the choice of sampling units, and in the formation of strata, etc. (United Nations 1964, p. 7).

See also SAMPLING UNITS.

## FRAME POPULATION

"The materials or devices which delimit, identify, and allow access to the elements of the target population" (Wright and Tsao 1983, p. 26).

"Is the set of persons for whom some enumeration can be made prior to the selection of the survey sample" (Groves 1989, p. 82).

## FRAME UNITS

See SAMPLING UNITS.

## IN-SCOPE UNIT

Sampling units that if properly classified would be part of the population of interest. They would be included in the frame if properly classified.

## LIST FRAMES

When the elements of the population have been numbered or otherwise identified, the population together with its identification system is called a list (Hansen, Hurwitz, and Madow 1953, Vol. II, p. 1).

Nongeographically defined units for drawing a sample (Hansen, Hurwitz, and Jabine 1963).

A list of all the sampling units in the population. This list provides the basis for the selection and identification of units in the sample (Sukhatme and Sukhatme 1970, p. 2).

## LISTING ERRORS

An error in a sample estimate that occurs due to a failure to find units which should be listed, failure to classify a unit as being within the scope of the list, listing a unit which is not within the scope of the list, or listing a unit more than once. As defined in this report, a type of coverage error that occurs in surveys in which the frame sampling unit and the ultimate sampling unit for the survey are different. Three basic listing errors are cited in this report: Area segment listing errors, household listing errors, and nonhousehold listing errors (this report, p. 39).

## LOCATION ERRORS

An error in a sample estimate that arises because of an incorrect association of reporting units with sampling units when the sampling units themselves are not uniquely or clearly defined or when they are difficult to locate. A type of rule-of-association error that is a source of coverage error (this report, p. 31).

## LONGITUDINAL COVERAGE ERROR

An error in a sample estimate that results from unaccounted-for changes in the sample population from the first to subsequent interviews. A type of temporal error that results in coverage error (this report, p. 36).

## LONGITUDINAL SURVEY

A survey in which data are gathered over an extended period of time (Bailey 1982, p. 34).

## NONCOVERAGE

"Noncoverage includes the problems of 'incomplete frames,' a term that seems to imply omissions in preparing the frame. But also refers to 'missed units,' omissions due to faulty execution of survey procedures" (Kish 1965, p. 528).

"Missing elements, also called noncoverage and incomplete frame" (Kish 1965, p. 56).

"Households missing from a telephone survey sampling frame" (Groves, et al. 1988, p. 4).

See also UNDERCOVERAGE.

## NONSAMPLING ERRORS

"An error in sample estimates which cannot be attributed to sampling fluctuations. Such errors may arise from many different sources such as defects in the frame, faulty demarcation of sample units, defects in the selection of sample units, mistakes in the collection of data due to personal variations or misunderstandings or bias or negligence or dishonesty on the part of the investigator or of the interviewer, mistakes at the stage of the processing of the data, etc." (Kendall and Buckland 1971).

"The error in an estimate arising at any stage in a survey from such sources as varying interpretation of questions by enumerators, unwillingness or inability of respondents to give correct answers, nonresponse, improper coverage, and other sources exclusive of sampling error. This definition includes all components of the Mean Square Error (MSE) except sampling variance" (U.S. Bureau of the Census (no date), p. 50).

## NONRESPONSE

See UNIT NONRESPONSE.

## OBSERVATION UNITS

"The units from which the observations are obtained. In interview surveys they are called respondents" (Kish 1965, p. 8).

## OBSERVATIONAL ERROR

"Observational errors are deviations of the answers of respondents from their true values on the measure" (Groves 1989, p. 11).

"Errors which are caused by obtaining and recording observations incorrect[e]ly" (Kish 1965, p. 520).

## OUT-OF-SCOPE ELEMENTS

Elements that if properly classified would not be part of the population of interest. If properly classified, they would be dropped from the frame (this report, p. 20).

See also CLASSIFICATION ERRORS.

## OUT-OF-SCOPE UNITS

Sampling units that if properly classified would not be part of the population of interest.

## OVERCOVERAGE

"Target population units [which] ... are duplicated or enumerated in error" (Hansen, Hurwitz, and Madow 1953, Vol. I, p. 84).

Errors leading to the inclusion of units which are not members of the target population (this report, p. 1).

Some members of the population are represented more than once (this report, p. 14).

## **PLANT**

See ESTABLISHMENT.

## **POPULATION OF INTEREST**

See TARGET POPULATION.

## **RANDOM-DIGIT DIALING**

"Random digit dialing (RDD) methods are based on the frame of all possible telephone numbers. The telephone number frame is commonly assembled by appending suffixes to area code-prefix combinations obtained from Bell Communications Research (BCR) for a fee." (Groves, et al. 1988, p. 81).

## **RECORDING ERRORS**

Error arising when information is correctly known but incorrectly recorded. As defined in this report, a type of nonsampling error that can cause coverage error (this report, p. 47).

## **RELEVANCE**

"Standards of relevance are concerned with the difference between the ideal goal of a survey and the statistics called for by the survey specifications" (Hansen, Hurwitz, and Pritzker 1967).

## **RELEVANCE ERROR**

See CONCEPTUAL ERROR.

## **RESPONDENT ERROR**

See RESPONDENT REPORTING ERROR.

## **RESPONDENT REPORTING ERROR**

In this report, respondent reporting error refers to all errors which occur during the interview process, whether they are caused by the interviewer, the respondent, vague concepts, faulty instructions, imprecise questions, or combined effects of several of these. A type of household listing error that can cause coverage error (this report, p. 43).

## **RULES OF ASSOCIATION**

Those rules which allow establishment of a linkage between a selection of listed units with known probabilities to a selection of reporting units with known probabilities (Hansen, Hurwitz, and Jabine 1963).

Delineate the relationship between sampling units and the final reporting unit (this report, p. 31).

Also known as rules of correspondence (Groves 1989, p. 99).

## **RULE-OF-ASSOCIATION ERROR**

An error in an estimate that results when the relationship between sampling units and the final reporting unit is delineated incorrectly. Rule-of-association errors have been classified into three basic types for purposes of this report: location errors, classification errors, and temporal errors. Rule-of-association errors are a source of coverage error (this report, p. 31).

## **SAMPLING FRAME**

See FRAME.

## **SAMPLING UNITS**

"These units must cover the whole of the population and they must not overlap, in the sense that every element in the population belongs to one and only one unit" (Cochran 1963, p. 7).

The population is subdivided into a finite number of distinct and identifiable units called sampling units (Sukhatme and Sukhatme 1970, p. 2).

"Contain the elements, and they are used for selecting elements into the sample" (Kish 1965, p. 8).

"A sampling unit is either a single sampling element or a collection of elements" (Bailey 1982, p. 85).

"Nonoverlapping collections of elements from the population that cover the entire population" (Scheaffer, Mendenhall, and Ott 1986, p. 21).

"The elements of the population from which we select the sample" (Hansen, Hurwitz and Madow 1953, Vol. II, p. 5).

## **TARGET POPULATION**

"The population about which information is wanted" (Cochran 1963, p. 6).

"The set of persons of finite size which will be studied" (Groves 1989, p. 82).

## **TEMPORAL ERRORS**

An error in an estimate that results when the frame or sample is not updated to represent the population of interest for the survey's reference period. A type of rule-of-association error that is a source of coverage error (this report, p. 36).

## **TYPE A NONINTERVIEW**

U.S. Bureau of the Census nomenclature used to indicate a noninterview of a household that is occupied by persons eligible for interview (this report, p. 34).

See also UNIT NONRESPONSE.

## TYPE B NONINTERVIEW

U.S. Bureau of the Census nomenclature used to indicate a noninterview of a household that is either unoccupied but could become occupied, or occupied by persons not eligible for interview (this report, p. 34).

## TYPE C NONINTERVIEW

U.S. Bureau of the Census nomenclature used to indicate a noninterview of a household because the sampling unit is ineligible for the sample (this report, p. 34).

## UNDERCOVERAGE

"... failure to include in the frame all units belonging to the defined population" (U.S. Bureau of the Census (no date), p. 48).

"The number of persons in telephone households who are not enumerated in sample households in a telephone survey" (Groves, et al. 1988, p. 4).

See also NONCOVERAGE.

## UNITS

See SAMPLING UNITS.

## UNIT NONRESPONSE

"Unit nonresponse occurs if a unit is selected for the sample and is eligible for the survey, but no response is obtained for the unit or the obtained response is unusable" (Madow, et al 1983, Vol. 1, p. 18).

"The failure to elicit responses for units of analysis in a population or sample because of various reasons such as absence from home, failure to return questionnaire, refusals, omission of one or more entries in a form, vacant houses, etc." (U.S. Bureau of the Census (no date), p. 50).

"We shall use the term nonresponse to refer to the failure to measure some of the units in the selected sample" (Cochran 1963, p. 355).

"In sample surveys, the failure to obtain information from a designated individual for any reason (death, absence, refusal to reply) is often called nonresponse" (Kendall and Buckland 1971).

"Nonresponse refers to many sources of failure to obtain observations (response, measurements) on some elements selected and designated for the sample" (Kish 1965, p. 532).

"Is an error of nonobservation. Nonresponse is the failure to obtain complete measurements on the survey sample" (Groves 1989, p. 133).

As cited in this report, a source of coverage error (p. 1).

## REFERENCES

- Adams, D. (1989), "Recommendation to Not Consider the Half-open Interval Approach to Sample Selection (Doc. #4.2-R-2)," Unpublished memorandum to Work Group 4, U.S. Bureau of the Census.
- Alexander, C. (1986), "The Present Consumer Expenditure Survey's Weighting Method," in *Population Controls and Weighting Sample Visits*, Washington, DC: U.S. Bureau of Labor Statistics.
- Anderson, D., Schoenberg, B., and Haerer, A. (1988), "Prevalence Surveys of Neurologic Disorders: Methodologic Implications of the Copiah County Study," *Journal of Clinical Epidemiology*, 41, 339-345.
- Armington, C., and Odle, M. (1981), "Sources of Employment Growth, 1978-1980," Mimeograph, Washington, DC: Brookings Institution, Business Microdata Project.
- Bailer, B. (1984), "The Quality of Survey Data," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 43-52.
- Bailey, K. (1982), *Methods of Social Research* (2nd ed.), New York: The Free Press.
- Beller, N. (1979), "Error Profile for Multiple-Frame Surveys," Economic Statistics and Cooperative Service Report 63, Washington, DC: U.S. Department of Agriculture. Also in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 221-222.
- Bernhardt, M., and Helfand, S. (1980), "Reconciliation of the Economic Censuses Results and Current Survey Programs," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 169-174.
- Biderman, A., and Cantor, D. (1984), "A Longitudinal Analysis of Bounding; Respondent Conditioning and Mobility as Sources of Panel Bias in the National Crime Survey," *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 708-713.
- Biemer, P. (1983), "Optimal Dual Frame Sample Design: Results of a Simulation Study," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 630-635.
- Birch, D.L. (1979), "The Job Generation Process," Mimeograph, Cambridge, MA: Institute of Technology, Program on Neighborhood and Regional Change.
- Blalock, H. (1968), "The Measurement Problem: The Gap Between the Languages of Theory and Research," in *Methodology in Social Research*, eds. H. Blalock and A. Blalock, New York: McGraw-Hill, pp. 1-27.
- Bosecker, R. (1984), "List vs. Area Overlap Determination: List Dominant and Frozen Domain Procedures," National Agricultural Statistics Service Staff Report, Washington, DC: U.S. Department of Agriculture.
- \_\_\_\_\_, and Clark, M. (1988), "Modifying the Weighted Estimator to Eliminate Screening Interviews in Residential Areas," National Agricultural Statistics Service Research Report, Washington, DC: U.S. Department of Agriculture.

- \_\_\_\_\_, and Kelly, W. (1975), "Summary of Results from Nebraska Concept Study," Statistical Reporting Service Staff Report, Washington, DC: U.S. Department of Agriculture.
- Casady, R., Nathan, G., and Sirken, M. (1985), "Alternative Dual System Network Estimators," *International Statistical Review*, 53, 183-197.
- Clogg, C., Massagli, M., and Eliason, S. (1986), "Population Undercount as an Issue in Social Research," in *Proceedings of the Second Annual Research Conference, U.S. Bureau of the Census*, pp. 335-343.
- Cochran, W.G. (1963), *Sampling Techniques*, New York: John Wiley and Sons, Inc.
- Cohen, S., Flyer, P., and Potter, D. (1987), "Sample Design of the Medical Expenditure Survey - Institutional Population Component," Paper presented at the Annual Meeting of the American Public Health Association, New Orleans, LA.
- Colledge, M. (1989), "Coverage and Classification Maintenance Issues in Economic Surveys," *Panel Surveys*, eds. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh, New York: John Wiley and Sons, pp. 80-107.
- Connor, J., Heeringa, S., and Jackson, J. (1985), "Measuring and Understanding Economic Change in Michigan," Mimeograph, University of Michigan, Institute for Social Research.
- Cook, P. (1985), "The Case of the Missing Victims: Gunshot Woundings in the National Crime Survey," *Journal of Quantitative Criminology*, 1, 91-102.
- Cotter, J., and Nealon, J. (1987), "Area Frame Design for Agricultural Surveys," National Agricultural Statistics Service Report, Washington, DC: U.S. Department of Agriculture.
- Coulter, R., and Mergerson, J. (1978), "An Application of a Record Linkage Theory in Constructing a List Sampling Frame," in *Proceedings of the Tenth Symposium on the Interface of Computer Science and Statistics*, pp. 416-420.
- Cowan, C., Breakey, W., and Fischer, P. (1988), "The Methodology of Counting the Homeless," in *Homelessness, Health and Human Needs*, Washington, DC: National Academy Press.
- Crank, K. (1979), "The Use of Current Partial Information to Adjust for Nonrespondents," Statistical Reporting Service Memorandum, Washington, DC: U.S. Department of Agriculture.
- Czaja, R., Snowden, C., and Casady, R. (1986), "Reporting Bias and Sampling Errors in a Survey of a Rare Population Using Multiplicity Counting Rules," *Journal of the American Statistical Association*, 81, 411-419.
- Dalenius, T. (1974), *Ends and Means of Total Survey Design*, Stockholm: University of Stockholm.
- \_\_\_\_\_. (1985), "Elements of Survey Sampling," Notes prepared for the Swedish Agency for Research Cooperation with Developing Countries (SAREC).
- Deming, W. (1960), *Sample Design in Business Research*, New York: John Wiley and Sons, Inc.



- \_\_\_\_\_. (1961), "Uncertainties in Statistical Data, and Their Relation to the Design and Management of Statistical Surveys and Experiments," *Bulletin of the International Statistical Institute*, 38, Part IV, 365-383.
- Fay, R. (1989), "An Analysis of Within-Household Undercoverage in the Current Population Survey," in *Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census*, pp. 156-175.
- \_\_\_\_\_, Passel, J., and Robinson, J. (1988), "The Coverage of Population in the 1980 Census," Publication PHC80-E4, Washington, DC: U.S. Government Printing Office.
- Fellegi, I., and Sunter, A. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183-1210. Also in *Record Linkage Techniques-1985*, eds. B. Kilss and W. Alvey, Washington, DC: Internal Revenue Service, pp. 51-78.
- Gleason, C., and Bosecker, R. (1978), "The Effect of Refusals and Inaccessibles in List Frame Estimates," Economic Statistics and Cooperative Service Report, Washington, DC: U.S. Department of Agriculture.
- Groves, R. (1989), *Survey Error and Survey Costs*, New York: John Wiley and Sons, Inc.
- \_\_\_\_\_, Biemer, P., Lyberg, L., Massey, J., Nicholls, W., and Waksberg, J. (eds.) (1988), *Telephone Survey Methodology*, New York: John Wiley and Sons, Inc.
- Grzesiak, T., and Tupek, A. (1987), "Measuring Employment of New Businesses in the Current Employment Statistics Survey," Paper presented at the International Roundtable on Business Survey Frames.
- Gurney, M., and Gonzalez, M. (1972), "Estimates for Samples from Frames When Some Units Have Multiple Listings," *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 283-288.
- Hainer, P. (1987), "A Brief and Qualitative Anthropological Study Exploring the Reasons for Census Coverage Error Among Low Income Black Households," Report submitted to U.S. Bureau of the Census.
- \_\_\_\_\_, Hines, C., Martin, E., and Shapiro, G. (1988), "Research on Improving Coverage in Household Surveys," *Proceedings of the Fourth Annual Research Conference, U.S. Bureau of the Census*, pp. 513-539.
- Hanczaryk, P., and Sullivan, J. (1980), "Evaluation of Coverage of the Administrative Records Frame for the 1977 Economic Censuses - Employer Segment," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 154-159.
- Hansen, M., Hurwitz, W., and Jabine, T. (1963), "The Use of Imperfect Lists for Probability Sampling at the U.S. Bureau of the Census," *Bulletin of the International Statistical Institute*, 40, 497-517.
- \_\_\_\_\_, \_\_\_\_\_, and Madow, W. (1953), *Sample Survey Methods and Theory* (Vol. I, Methods and Applications), New York: John Wiley and Sons, Inc.
- \_\_\_\_\_, \_\_\_\_\_, and \_\_\_\_\_ (1953), *Sample Survey Methods and Theory* (Vol. II, Theory), New York: John Wiley and Sons, Inc.

- \_\_\_\_\_, \_\_\_\_\_, and Pritzker, H. (1967), "Standardization of Procedures for Data, Measurement Errors and Statistical Standards in the Bureau of the Census," *Bulletin of the International Statistical Institute*, 42, Part I, 49-64.
- Hanson, R. (1978), "The Current Population Survey: Design and Methodology," Technical Paper No. 40, Washington, DC: U.S. Government Printing Office.
- Hartley, H.O. (1962), "Multiple Frame Surveys," *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 203-206.
- Harwood, A. (1970), "Participant Observation and Census Data in Urban Research," Paper presented at the Annual Meeting of the American Anthropological Association, San Diego, CA.
- Hauber, F., Bruininks, R., Hill, R., Lakin, K., and White, C. (1984), "National Census of Residential Facilities: Fiscal Year 1982," Report, University of Minnesota, Department of Educational Psychology.
- Hawkes, W., Jr. (1985), "Census Data Quality - A User's View," *Proceedings of the First Annual Research Conference, U.S. Bureau of the Census*, pp. 177-192.
- Hill, G., and Rockwell, D. (1977), "Associating a Reporting Unit with a List Frame Sampling Unit," Internal memorandum, U.S. Department of Agriculture.
- Hirschberg, D., Yuskavage, R., and Scheuren, F. (1977), "The Impact on Personal and Family Income of Adjusting the CPS for Undercoverage," *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 70-80.
- Jacobs, C. (1986), "Interim Evaluation of Listing Process Audit," Unpublished memorandum to Housing Working Group, U.S. Bureau of Labor Statistics.
- Jean, A., and McArthur, E. (1984), "Some Data Collection Issues for Panel Surveys with Application to the SIPP," Survey of Income and Program Participation Working Paper Series No. 8407, Washington, DC: U.S. Bureau of the Census.
- Jessen, R. (1978), *Statistical Survey Techniques*, New York: John Wiley and Sons, Inc.
- Johnston, D., and Wetzell, J. (1969), "Effect of the Census Undercount on Labor Force Estimates," *Monthly Labor Review*, March, 3-13.
- Joncas, M. (1985), "Cluster Listing Check Program for the Redesign LFS Sample," Unpublished report, Ottawa: Statistics Canada.
- Kalton, G., and Anderson, D. (1986), "Sampling Rare Populations," *Journal of the Royal Statistical Society - A*, 149 (Pt. 1), 65-82.
- \_\_\_\_\_, and Lepkowski, J. (1985), "Following Rules in SIPP," *Journal of Economic and Social Measurement*, 13, 319-328.
- Kendall, M.G., and Buckland, W.R. (1971), *A Dictionary of Statistical Terms* (3rd ed.), International Statistical Institute.
- King, K. (1988), "SIPP: Monitoring the Rates of Ineligible Households," Internal memorandum, U.S. Bureau of the Census, February 24.

- \_\_\_\_\_, Petroni, R., and Singh, R. (1987), "Quality Profile for the Survey of Income and Program Participation," SIPP Working Paper Series 8708, Washington, DC: U.S. Bureau of the Census.
- Kish, L. (1965), *Survey Sampling*, New York: John Wiley and Sons, Inc.
- Konschnik, C. (1987), "Summary of the Results of the Area Sample Recheck for the Period Aug. 1986-Oct. 1986," Internal memorandum, U.S. Bureau of the Census.
- Lepkowski, J. (1988), "Telephone Sampling Methods in the United States," in *Telephone Survey Methodology*, eds. R. Groves, et al., New York: John Wiley and Sons, Inc., pp. 73-98.
- \_\_\_\_\_, and Groves, R. (1986), "A Mean Squared Error Model for Dual Frame, Mixed Mode Survey Design," *Journal of the American Statistical Association*, 81, 930-937.
- Lessler, L. (1980), "Errors Associated With the Frame," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 125-130.
- Linebarger, J. (1975), "New Construction Time-Lag Study," Internal memorandum to C. Bostrum dated August 22, U.S. Bureau of the Census.
- Madow, W., Nisselson, H., and Olkin, I., (eds.) (1983), *Incomplete Data in Sample Surveys*, 3 Vols., New York: Academic Press.
- Manton, K. (1988), "A Longitudinal Study of Functional Change and Mortality in the United States," *Journal of Gerontology*, 43, 5153-5161.
- Marks, E., and Nisselson, H. (1977), "Problems of Nonsampling Error in the Survey of Income and Education: Coverage Evaluation," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 414-417.
- \_\_\_\_\_, Seltzer, W., and Krotki, K. (1974), *Population Growth Estimation: A Handbook of Vital Statistics Measurement*, New York: The Population Council.
- Martin, E. (1981), "A Twist on the Heisenberg Principle: Or, How Crime Affects Its Measurement," *Social Indicators Research*, 9, 197-223.
- Matthews, R. (1988), "Screening Residential Tracts for Agricultural Activity," National Agricultural Statistics Service Staff Report, Washington, DC: U.S. Department of Agriculture.
- McArthur, E., and Short, K. (1986), "Measurement of Attrition from SIPP Through the Fifth Wave of the 1984 Panel," Internal memorandum, U.S. Bureau of the Census.
- McDonald, R. (1984), "The 'Underground Economy' and BLS Statistical Data," *Monthly Labor Review*, January, 4-18.
- McGowan, H. (1982), "Telephone Ownership in the National Crime Survey," Unpublished memorandum, U.S. Bureau of the Census.
- Montie, I.C., and MacKenzie, W. (1978), "Open-Ended Segments: Variation on Area Segmenting and List Frame Supplementation," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 233-237.

- Montie, I.C. and Schwanz, D.J. (1977), "Coverage Improvement in the Annual Housing Survey," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 163-172.
- Moser, C.A., and Kalton, G. (1971), *Survey Methods in Social Investigation*, (2nd ed.), New York: Basic Books, Inc.
- Nealon, J. (1984), "Review of the Multiple Frame and Area Frame Estimator," National Agricultural Statistical Service Staff Report, Washington, DC: U.S. Department of Agriculture.
- Newbrough, U. (1988), "CPS Reinterview Quality Control Results for 1987," Internal memorandum, U.S. Bureau of the Census.
- Parsons, V., and Casady, R. (1986), "Variance Estimation and the Redesigned National Health Interview Survey," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 406-411.
- Pennie, K. (1990), "Coverage Comparisons Between the 1980 Census and the Current Population Survey," Internal memorandum to P. Waite, U.S. Bureau of the Census, in draft.
- Pollard, A., Usef, F. and Pollard, G. (1974), *Demographic Techniques*, Rushcutters Bay, Australia: Pergamon Press Pty. Ltd.
- Potter, D., Cohen, S., and Mueller, C. (1987), "The 1986 Inventory of Long-term Care Places as a Sampling Frame," Paper presented at the Annual Meeting of the American Statistical Association, San Francisco, CA.
- Research Triangle Institute (RTI) (1981), "Complementing the Survey of Hospitals and Other Health Institutions (1980 Complement Survey)," Project report, RTI Contract No. 255 U-1913, Research Triangle Park, NC: Author.
- Scheaffer, R., Mendenhall, W., and Ott, W. (1986) *Elementary Survey Sampling* (3rd ed.), Boston: Duxbury Press.
- Scheuren, F., and Oh, H.L. (1985), "Fiddling Around with Nonmatches and Mismatches," in *Record Linkage Techniques-1985*, eds. B. Kilss and W. Alvey, Washington, DC: U.S. Internal Revenue Service, pp. 79-88.
- Schreiner, I. (1987), "CPS Reinterview Quality Control Results for 1986," Internal memorandum, U.S. Bureau of the Census.
- Schwanz, D. (1988a), "1985 Type-A Unable-to-Locate Rates for the AHS National Unit Samples," Internal memorandum, U.S. Bureau of the Census.
- \_\_\_\_ (1988b), "Mobile Home New Construction for 1985 AHS - National," Internal memorandum to Ed Montfort dated February 3, U.S. Bureau of the Census.
- Science Applications International Corporation (SAIC) (1985), "Evaluation of the Quality and Utility of the National Master Facility Inventory," Final report to the Division of Health Care Statistics, National Center for Health Statistics Contract 282-83-2114, Vienna, VA: Author.

- Shapiro, G. (1979), "Coverage Comparisons Between the Census and Current Population Survey (CPS)," Internal memorandum to Charles D. Jones, U.S. Bureau of the Census.
- \_\_\_\_ (1986), "Second Set of Experimental Interviews," Internal memorandum, U.S. Bureau of the Census, July 16.
- \_\_\_\_, and Kostanich, D. (1988), "High Response Error and Poor Coverage Are Severely Hurting the Value of Household Surveys," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 443-448.
- Shimizu, I. (1983), "Identifying and Obtaining the Yellow Pages for a National Area Sample," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 558-562.
- \_\_\_\_ (1986), "The 1985 National Nursing Home Survey Design," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 516-520.
- Singh, R. (1989), "SIPP 85: Household Coverage," Internal memorandum to G. Shapiro, U.S. Bureau of the Census, June 22.
- Sirken, M. (1970), "Household Surveys with Multiplicity," *Journal of the American Statistical Association*, 65, 257-266.
- \_\_\_\_ (1983), "Handling Missing Data by Network Sampling," in *Incomplete Data in Sample Surveys*, eds. W. Madow, H. Nisselson, and I. Olkin, Vol. 2, New York: Academic Press, pp. 81-90.
- \_\_\_\_, and Levy, P. (1974), "Multiplicity Estimation of Proportions Based on Ratios of Random Variables," *Journal of the American Statistical Association*, 19, 68-74.
- \_\_\_\_ and Royston, P. (1976), "Effect of Selected Survey Design Factors on the Registered Deaths Reported in a Single-time Retrospective Household Survey," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 773-777.
- Spillman, B. (1989), Internal memorandum to A. Swell, National Center for Health Services, June 12.
- Statt, R., Vacca, E., Wolters, C., and Hernandez, R. (1981), "Problems Associated with Using Building Permits as a Frame of Post-Census Construction: Permit Lag and ED Identification," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 226-231.
- Strahan, G. (1987), "Nursing Home Characteristics: Preliminary Data from the 1985 National Nursing Home Survey," *Advance Data from Vital and Health Statistics*, No. 131, Publication PHS-87-1250, Washington, DC: U.S. Government Printing Office.
- Sukhatme, P.V., and Sukhatme, B.V. (1970), *Sampling Theory of Surveys with Applications* (2nd rev. ed.), Ames, IA: Iowa State University Press.
- Teitz, M., Glasmeier, A., and Svensson, D. (1981), "Small Business and Employment Growth in California," Working Paper No. 348, University of California at Berkeley, Institute of Urban and Regional Development.

- Thomas, A. (1986), "BLS Establishment Estimates Revised to March 1985 Benchmarks," Washington, DC: U.S. Bureau of Labor Statistics.
- Thornberry, O., and Massey, J. (1988), "Trends in United States Telephone Coverage Across Time and Subgroups," in *Telephone Survey Methodology*, eds. R. Groves, et al., New York: John Wiley and Sons, Inc., pp. 25-51.
- Tortora, R. (1987), "Quantifying Nonsampling Errors and Bias," *Journal of Official Statistics* (Statistics Sweden), 3, 339-342.
- United Nations (1964), *Recommendations for Preparation of Sampling Survey Reports* (Provisional Issue), Series C, No. 1 Rev. 2, New York.
- \_\_\_\_ (1982), *Nonsampling Errors in Household Surveys: Sources, Assessment and Control*, New York: Author.
- U.S. Bureau of Labor Statistics (1985), *Employment and Earnings*, 32 (12).
- \_\_\_\_ (1986), *Consumer Expenditure Survey*, Washington, DC: U.S. Government Printing Office.
- \_\_\_\_ (1988), *BLS Handbook of Methods*, Bulletin 2285, Washington, DC: U.S. Government Printing Office.
- \_\_\_\_ (1989), *Employment and Earnings*, 36 (12).
- U.S. Bureau of the Census (no date), *Course on Nonsampling Errors*, Lectures 1-9, International Statistics Program Center, Washington, DC.
- \_\_\_\_ (1968), "The Current Population Survey Reinterview Program, January 1961 through December 1966," Technical Paper 19, Washington, DC: U.S. Government Printing Office.
- \_\_\_\_ (1971), "The Annual Survey of Manufactures: A Report on Methodology," Technical Paper 24, Washington, DC: U.S. Government Printing Office.
- \_\_\_\_ (1973), "The Coverage of Housing in the 1970 Census," Report PHC(E)-5, Washington, DC: U.S. Government Printing Office.
- \_\_\_\_ (1984), "Census of Agriculture, Coverage Evaluation," Vol. 2, Pt. 2, Publication AC82-SS-2, Washington, DC: U.S. Government Printing Office.
- \_\_\_\_ (1986), "Appendix A: Source and Reliability Statement for the Long-term Care Survey," in *National Long-term Care Survey and National Survey of Informal Caregivers, 1982 Report on Methods and Procedures Used in the Survey, Part 1, Documentation*, Springfield, VA: U.S. National Technical Information Services, Order No. 86-161783, p. A-3.
- \_\_\_\_ (1987), "Programs to Improve Coverage in the 1980 Census," Report PHC80-E3, Washington, DC: U.S. Government Printing Office.
- \_\_\_\_ (1989), *Current Construction Reports, Housing Starts*, C20-89-4, Housing Starts Compilation, Washington, DC: U.S. Government Printing Office.

U.S. Department of Commerce (1978a), "An Error Profile: Employment as Measured by the Current Population Survey," Statistical Policy Working Paper 3, Washington, DC: U.S. Government Printing Office.

\_\_\_\_ (1978b), "Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics," Statistical Policy Working Paper 4, Washington, DC: U.S. Government Printing Office.

\_\_\_\_ (1980), "Report on Statistical Uses of Administrative Records," Statistical Policy Working Paper 6, Washington, DC: U.S. Government Printing Office.

U.S. Executive Office of the President, Office of Management and Budget (1987), Standard Industrial Classification Manual: 1987, Order No. PB87-1000012, Springfield, VA: National Technical Information Service.

U.S. National Center for Health Statistics (1965), "Development and Maintenance of a National Inventory of Hospitals and Institutions," *Vital and Health Statistics*, Series 1, No. 6 (PHS Publication Number 1000), Washington, DC: U.S. Government Printing Office.

\_\_\_\_ (1968), "The Agency Reporting System for Maintaining the National Inventory of Hospitals and Institutions," *Vital and Health Statistics*, Series 1, No. 6 (PHS Publication Number 1000), Washington, DC: U.S. Government Printing Office.

\_\_\_\_ (1983), "Nursing and Related Care Homes as Reported from the 1980 National Master Facility Inventory Survey," by A. Sirrocco, *Vital and Health Statistics*, Series 14, No. 29 (PHS Publication Number 84-1824), Washington, DC: U.S. Government Printing Office.

\_\_\_\_ (1986), "Nursing and Related Care Homes as Reported from the 1982 National Master Facility Inventory Survey," by D. Roper, *Vital and Health Statistics*, Series 14, No. 32 (PHS Publication Number 86-1827), Washington, DC: U.S. Government Printing Office.

\_\_\_\_ (1987), "Public Use Data Tape Documentation: The 1986 Inventory of Long-term Care Places," NTIS Publication 88-110614, Springfield, VA: National Technical Information Service.

U.S. Office of Management and Budget (1986), "Federal Longitudinal Surveys," Statistical Policy Working Paper 13, Washington, DC: U.S. Government Printing Office.

Valentine, C., and Valentine, B. (1971), "Summary of Missing Men - A Comparative Methodological Study of Underenumeration and Related Problems," Unpublished paper, U.S. Bureau of the Census.

Vogel, F., and Bosecker, R. (1974), "Multiple Frame Livestock Surveys, A Comparison of Area and List Sampling," National Agricultural Statistics Service Staff Report, Washington, DC: U.S. Department of Agriculture.

\_\_\_\_, and Rockwell, D. (1977), "Fiddling with Area Frame Information in List Development and Maintenance," Washington, DC: U.S. Department of Agriculture.

Waite, P.J. (1989), "Listing Accuracy in HIS Blocks," Internal memorandum to E. Davey and S.D. Matchett dated October 26, U.S. Bureau of the Census.

Waksberg, J. (1978), "Sampling Methods for Random Digit Dialing," *Journal of the American Statistical Association*, 73, 40-46.

Williams, L., and Chakrabarty, R. (1983), "The Michigan State Random Digit Dialing Survey of Sportsmen and Wildlife Associated Recreation," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 648-653.

Wright, T., and Tsao, H. (1983), "A Frame on Frames: An Annotated Bibliography" in *Statistical Methods and the Improvement of Data Quality*, ed. T. Wright, New York: Academic Press, pp. 25-72.



**Reports Available in the  
Statistical Policy  
Working Paper Series**

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document Sales, PB86-211539/AS)
3. An Error Profile: Employment as Measured by the Current Population Survey (NTIS Document Sales PB86-214269/AS)
4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)
5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)
6. Report on Statistical Uses of Administrative Records (NTIS Document Sales, PB86-214285/AS)
7. An Interagency Review of Time-Series Revision Policies (NTIS Document Sales, PB86-232451/AS)
8. Statistical Interagency Agreements (NTIS Document Sales, PB86-230570/AS)
9. Contracting for Surveys (NTIS Document Sales, PB83-233148)
10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)
11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276)
12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)
13. Federal Longitudinal Surveys (NTIS Document Sales, PB86-139730)
14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)
15. Quality in Establishment Surveys (NTIS Document Sales, PB88-232921)
16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB90-205238)
17. Survey Coverage (NTIS Document Sales, PB90-205246)
18. Data Editing in Federal Statistical Agencies (NTIS Document Sales, PB90-205253)
19. Computer Assisted Survey Information Collection (NTIS Document Sales, PB90-205261)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703) 487-4650